

MICROECONOMICS

SECOND EDITION

HUGH GRAVELLE
AND
RAY REES



LONGMAN

LONDON AND NEW YORK

Contents

Preface to the second edition	xi
1. The nature and scope of microeconomics	1
A. Concepts and methods	1
B. The economic and social framework	9
References and further reading	11
2. Optimization	12
A. The structure of an optimization problem	12
B. Solutions: questions and concepts	16
C. Existence of solutions	25
D. Local and global optima	27
E. Uniqueness of solutions	30
F. Interior and boundary optima	31
G. Location of the optimum: the method of Lagrange	33
H. Concave programming and the Kuhn-Tucker conditions*	42
I. Second-order conditions and comparative statics*	53
J. The Envelope Theorem	64
K. Conclusions	66
Notes	66
References and further reading	67

* Denotes sections containing material which can be omitted on a first reading either because it is more advanced or uses concepts developed later in the text.

Group UK Limited
House, Burnt Mill,
Essex CM20 2JE, England
associated Companies throughout the world.

in the United States of America
Longman Publishing, New York

Longman Group UK Limited 1981, 1992

reserved; no part of this publication may be
d, stored in a retrieval system, or transmitted
m or by any means, electronic, mechanical,
/ing, recording, or otherwise without either the
ten permission of the Publishers or a licence
restricted copying in the United Kingdom issued
Copyright Licensing Agency Ltd,
Ham Court Road, London W1P 9HE.

shed 1981
ition 1992

023866 PPR

ary Cataloguing-in-Publication Data

e record for this book is
om the British Library

Congress Cataloging in Publication Data
ugh.

economics/Hugh Gravelle and Ray Rees.—2nd ed.
cm.

s bibliographical references and index.

1-582-02386-6 (pbk.)

roeconomics. I. Rees, Ray, 1943— II. Title.
G786 1992

20 92-7641
CIP

5 in 10/12pt Monotype Times

/ Longman Group (FE) Limited
ingapore

3. The theory of the consumer	68
A. The preference ordering	68
B. The feasible set	80
C. The consumption decision	82
D. The comparative statics of consumer behaviour	87
E. Offer curves and net demand curves	95
Appendix 1: The lexicographic ordering*	99
Appendix 2: Existence of a utility function*	101
References and further reading	104
4. Consumer theory: duality	105
A. The expenditure function	105
B. The indirect utility function, Roy's identity and the Slutsky equation	111
C. Measuring the benefits of price changes	116
D. Composite commodities, separability and homotheticity*	124
E. Aggregation*	129
Notes	132
References and further reading	132
5. Further models of consumer behaviour*	133
A. Revealed preference	133
B. Consumption technology	139
C. The consumer as a labour supplier	149
D. Consumption and the allocation of time	157
References and further reading	164
6. The firm	166
A. Introduction	166
B. The nature of the firm	166
C. Critique of the classical theory of the firm	170
D. Issues in the theory of the firm	172
E. Conclusions	179
References and further reading	179
7. Production	180
A. The production function	180
B. Variations in scale	186
C. Variations in input proportions	190
D. The multi-product case	192
References and further reading	195

8. Cost	197
A. Introduction	197
B. Long-run cost minimization	200
C. Short-run cost minimization	212
D. Cost minimization with several plants	222
E. Multi-product cost functions*	226
References and further reading	229
9. Supply	230
A. Long-run profit maximization	231
B. Short-run profit maximization	235
C. The multi-product firm	238
D. The profit function and comparative statics*	241
References and further reading	246
10. The theory of competitive markets	247
A. Short-run equilibrium	248
B. Stability of equilibrium	253
C. Long-run equilibrium	262
Conclusions	267
References and further reading	268
11. Monopoly	269
A. Introduction	269
B. Price and output determination under monopoly	270
C. Price discrimination	274
D. Entry	285
Notes	297
References and further reading	297
12. Oligopoly	298
A. Introduction	298
B. One-shot games	300
C. Oligopoly as a repeated game	318
D. Credible threats*	326
E. Conclusion	337
Notes	338
References and further reading	339

13. Alternative theories of the firm	340
A. Introduction	340
B. The entrepreneurial firm	340
C. Agency theory and the separation of ownership from control*	346
D. Labour managed firms	354
E. Conclusions	358
References and further reading	359
14. Input markets and bargaining	360
A. Demand for inputs	360
B. Monopsony	368
C. Unions as monopoly input suppliers	373
D. Bilateral monopoly	377
E. Comparative bargaining games*	380
F. Bargaining as a non-cooperative game*	388
G. Delay and disagreement in bargaining*	395
References and further reading	404
15. Investment and consumption over time	406
A. Introduction	406
B. Optimal consumption over time	406
C. The optimal investment decision	411
D. Capital market equilibrium	420
E. Extensions: many periods; adjustment costs	426
References and further reading	436
16. General equilibrium	438
A. Introduction	438
B. Walrasian equilibrium of a competitive economy	439
C. Existence of Walrasian equilibrium	442
D. Stability of Walrasian equilibrium	449
E. Edgeworth exchange theory	454
F. Exchange, equilibrium and the core	458
Appendix	468
References and further reading	474
17. Welfare economics	475
A. Introduction	475
B. Pareto efficient resource allocation	475

	Contents	ix
C. Welfare functions and the Pareto criterion		485
D. Pareto efficiency and competitive markets		490
E. Distribution and markets		496
F. Arrow's impossibility theorem		502
References and further reading		511
18. Market failure and the second best		512
A. The causes of market failure		512
B. Instances of market failure		516
C. The theory of second best		534
D. Government action and government failure		539
References and further reading		544
19. Choice under uncertainty		546
A. Introduction		546
B. A formalization of 'uncertainty'		547
C. Choice under uncertainty		549
D. Properties of the utility function		557
E. Risk aversion and indifference curves		569
F. Measures of risk		576
References and further reading		584
20. Exchange under uncertainty		586
A. Introduction		586
B. The insurance decision		586
C. Risk spreading: the Arrow-Lind Theorem		594
D. Risk pooling and diversification		599
E. Exchange of state contingent income claims		606
F. Stock markets		617
G. State contingent commodities		628
H. Valuation of price changes under uncertainty		635
References and further reading		641
21. Production under uncertainty		643
A. Introduction		643
B. Competitive firm under uncertainty		643
C. Production with futures markets		656
D. Constrained efficiency		659
References and further reading		669

22. Asymmetric information and incomplete markets	671
A. Introduction	671
B. Asymmetric information in insurance markets: adverse conditions	671
C. Asymmetric information in insurance markets: moral hazards	681
D. Principal-agent theory: hidden information*	689
E. Principal-agent theory: hidden action*	701
F. Signalling	714
G. Incomplete insurance markets	719
H. Incomplete stock markets	725
References and further reading	734
 Index	 736

Preface to the second edition

The aim of the second edition is the same as the first: to provide a comprehensive exposition of modern microeconomic theory, beginning at the intermediate level and ending at a level appropriate for graduate students. It can be used for intermediate courses, for specialized senior undergraduate courses in advanced microeconomic theory and mathematical economics, and for first courses in graduate microeconomics.

This edition is a very extensive revision of the original. Almost every chapter contains a substantial amount of new material, about half the chapters have been completely rewritten, and we have added two new chapters on uncertainty and asymmetric information. Broadly, the revisions and additions take account of the important developments in microeconomic theory over the last ten years, and increase the book's suitability for use by advanced undergraduate and graduate students.

The extent of the coverage and the range of levels has a number of advantages. Students on advanced courses find it useful to have intermediate-level material for revision. Students beginning their intermediate course can regard this book as a companion for their entire period of study of microeconomics, including, if they go on, graduate studies. The range of microeconomic models presented provides an instructor with a great deal of flexibility in choosing the content and level of courses.

The broad coverage evolved as we incorporated material that seemed to us to be both interesting and important because it showed students of intermediate microeconomics that when the restrictive assumptions underlying the basic models are relaxed, microeconomic theory not only survives, it flourishes. We are pleased that this view has passed the market test, hence this new edition.

A second reason for the length of the book is the importance we place on exposition. The aim has always been to give the student an understanding of the economic content of a model, rather than just statements of its mathematics. We have taken whatever space we felt was necessary to achieve this purpose, using words and diagrams, as well as mathematics.

The mathematical prerequisites are:

1. Sets: set inclusion; union, intersection and partitions of sets; the meaning of necessary and sufficient conditions and of \Rightarrow and \Leftrightarrow notations.

2. Linear algebra: vectors as arrays of numbers and points in the appropriate coordinate space; addition, subtraction and inner products of vectors; convex combinations; matrices.
3. Calculus: functions, limits and continuity; partial differentiation; unconstrained maximization of functions of several variables.
4. Probability: probability laws; conditional probability; expected values; variance; covariance.

No particular facility with these topics is assumed. This will develop as the reader applies them repeatedly to the study of microeconomics.

Methods of constrained optimization are the foundation for much of microeconomics at this level. The reader will find it useful to work through Chapter 2 which gives a reasonably simple exposition of optimization theory using only the first three prerequisites listed above.

As in the first edition, the problems are an important part of the book. To include the answers would have lengthened the text considerably or led to the omission of much interesting material. We have therefore produced a workbook in which we provide fairly full answers to the problems, as well as setting additional ones. In several parts of the book, we leave as an exercise for the reader an extension or variation of a model, or some details of the analysis, to save space or not to interrupt the main line of exposition. The fact that these exercises are covered in the workbook allows us to do this with a clear conscience. The workbook can be regarded as a complementary textbook, with more emphasis on the technical aspects of problem-solving.

For helpful comments on the first edition or on drafts of the second we would like to thank Dick Allard, Patricia Apps, Antonie Bauer, Richard Cornes, Brian Ferguson, Glen Jones, Mike Hoy, Maria Iacovou, Gerhard Illing, Ngo Van Long, Elizabeth Strange and many generations of students. We are grateful to Chris Harrison of Longman for editorial stoicism and good-natured cooperation, and to Joy Cash for seeing the edition through a complicated and tightly time-constrained production process so efficiently. Whilst working on this edition, Ray Rees visited the CES, University of Munich; CORE; Nuffield College, Oxford; and the University of Warwick; and Hugh Gravelle visited Wolfson College, Oxford and the universities of Harvard, Manchester, and Stanford. We are grateful for hospitality and conversations there with colleagues.

H.S.E.G.

R.R.

CHAPTER 1

The nature and scope of microeconomics

A. Concepts and methods

Microeconomics is a set of theories with one aim: to help us gain an understanding of the process by which scarce resources are allocated among alternative uses, and of the role of prices and markets in this process. In its purest form, it is a philosophical inquiry into the processes of resource allocation. However, with understanding usually comes the ability to predict and to control, and this has been the case in microeconomics. The concepts and relationships economists have developed in their attempt to understand the workings of the economy provide the basis for the design of policies by governments wishing to influence the outcome of this process, or alternatively for a critique of the actions governments might take. Through the development of 'operations research', 'management science' and 'business economics', concepts from microeconomics have been applied to assist rational decision-taking in business.

A good way of providing an introductory overview of microeconomics is to set out its basic elements.

1. Goods and services or commodities

These are the central objects of economic activity, since 'economic activity' consists of the production and exchange of commodities. We distinguish commodities from each other by one or more of three characteristics: their *physical* nature and attributes, which determine the way in which they meet the needs of consumers and producers; the *location* at which they are made available; and the *date* at which they are made available. For example, coal and crude oil are physically different commodities, as are the services of a hairdresser and those of an accountant (though in each case, the broad category of resource from which the commodities derive – 'land' in one case and 'labour' in the other – is the same). Equally important is the fact that crude oil in Dubai available tomorrow is a different commodity from crude oil available tomorrow at a refinery in Western Europe; while coal in London today is a different commodity from coal in London this time next year. The basis of the distinction between commodities is that they cannot be regarded as perfect

substitutes in production or consumption – a businessman who goes along to his accountant for advice on a tax problem would not be just as happy to be offered a haircut instead.

In most of microeconomics we usually assume a *finite* set of possible physical bundles of attributes, a finite set of possible locations – we do not regard geographical space as continuous, but rather divided up into small areas – and a finite set of dates. We do not regard calendar time as continuous, but rather divided up into equal discrete time intervals, and moreover not as extending indefinitely far into the future, but instead we assume some definite, though possibly very distant, time horizon. These assumptions ensure that there is a finite number of commodities to be taken into account in our theories. Alternatively, we could assume a *continuum* of commodities: given any one commodity, we could always define another which is as close as we like to the first in attributes, location and time. Moreover, this commodity continuum need not be bounded – we could picture commodities as points on a line which stretches to infinity, since we could always define commodities available later in time. The methods of analysis required for an economy with such a commodity continuum differ sharply from those conventionally used in economics. Since the assumptions required to establish a finite set of commodities do not seem to do serious injustice to reality, while considerably simplifying the analysis, we gladly adopt them.

2. Prices

Associated with each commodity is a price, which may be expressed in one of two ways. First, we may choose one commodity in the economy as a *numeraire*, i.e. as the commodity in terms of which all prices are to be expressed. For example, suppose we choose gold. Then the price of each commodity is the number of units of gold which exchange for one unit of that commodity. The price of gold of course is 1. In general, we are free to choose any commodity as numeraire, so that prices could just as well be expressed in terms of the number of units of some kind of labour service which exchange for one unit of each other commodity. It might be argued that in reality different commodities may have different degrees of suitability for use in market transactions. Commodities which are not easily divisible, and which are bulky and subject to physical decay, will tend not to be used as a means of payment. However, it is important to note that a numeraire is not intended to represent a *means of exchange*, or 'money', in this sense. We are simply using it as a *unit of account*, or a *unit of measurement* for prices in the economy, and *nothing need be implied about the mechanism by which transactions actually take place*. Given the choice of numeraire, prices are effectively *commodity rates of exchange* – they express the rate at which the numeraire exchanges for each other commodity. They have the dimension (units of the numeraire/units of the commodity). They are therefore not independent of the units in which we measure commodities. For example, if we double the unit in which we measure each commodity *except* for the numeraire we would have to double prices (explain why).

The second way in which prices might be expressed does not involve a numeraire. Instead, we suppose there to be some unit of account which is not a quantity of some physical commodity, but an abstract unit used in making bookkeeping entries. If one unit of a commodity is sold, the account is credited with a certain number of units of account, while if the commodity is bought, the same number of units is debited from the account.

The price of the commodity is then the number of units debited or credited per unit of the commodity. We find it useful to give this unit of account a name, and so we could call it the £ sterling, or the \$US, for example. If different accounts are kept in different units, then rates of exchange between units of account must be established before transfers from one account to another can be made. Clearly, there is no physical substance corresponding to the unit of account, say the £ sterling. A cheque made out for £x is an instruction to credit one account and debit another, i.e. to transfer x units of account between accounts. Notes and coin have no intrinsic worth (until perhaps they cease to be used in exchange and acquire intrinsic worth – become commodities themselves – to numismatists), but are simply tokens representing numbers of units of account which are passed around directly and form part (usually a relatively small part) of the credit side of one's accounts.

The somewhat abstract way of expressing prices in terms of units of account therefore corresponds to the way prices are expressed in reality, and has come about because of the development of the modern banking system. There is, however, a straightforward correspondence between prices expressed in terms of units of account and prices expressed as commodity rates of exchange. Thus, suppose we have the set of prices expressed in £ sterling: $p_1, p_2 \dots p_n$. Then, by taking any one such price, say the n th, and forming the n ratios:

$$r_1 = p_1/p_n; \quad r_2 = p_2/p_n; \quad \dots r_n = p_n/p_n = 1 \quad [A.1]$$

we can interpret each $r_j, j = 1, 2, \dots, n$, as the number of units of commodity n which will exchange for one unit of commodity j , i.e. as commodity rates of exchange with n as the numeraire. Each r_j will be in dimensions (units of good n /units of good j) as we can see from:

$$\begin{aligned} p_j/p_n &= (\text{£/units of good } j \div \text{£/units of good } n) \\ &= (\text{units of good } n/\text{units of good } j), \quad j = 1, 2, \dots, n \end{aligned} \quad [A.2]$$

Thus, each r_j is the number of units of good n we could buy if we sold a unit of good j and spent the proceeds (p_j units of account) on good n .

3. Markets

The everyday notion of a market is as a specific place where certain types of commodities are bought and sold, for example a cattle market, or a fruit and vegetable market. The concept of a market in economics is much more general than this: a market exists whenever two or more individuals are prepared to enter into an exchange transaction, regardless of time or place. Thus, if two poachers meet in the middle of a forest in the dead of night, one with a catch of salmon and the other with a bag of pheasants, and they decide to negotiate an exchange of fish for fowl, we would say that a market exists. The word 'market' denotes exchange. The central problem in microeconomics is the analysis of how markets operate, since we view the process of resource allocation as a market process – a resource allocation is brought about by the workings of markets. For every commodity, therefore, a market does or will exist, and something which cannot be exchanged upon a market is not, from the point of view of microeconomics, a commodity.

It is important to distinguish between *forward* and *spot* markets. On a spot market, an agreement is made under which delivery of a commodity is completed within the current period; on a forward market, delivery will be made at some future period. (Some markets may do both, e.g. the market in leasehold accommodation, where what may be sold is a flow of housing services over possibly a very large number of years). We could envisage an economy in which at a given point in time there exists a market for every commodity, which means there is a complete system of spot and forward markets. In such an economy, contracts would be entered into for all future exchanges of commodities as well as for all current exchanges, and so market activity could cease entirely after the first period: the rest of the time would be spent simply fulfilling the contracts already concluded. Real economies, of course, do not possess such complete market systems. In any one period, markets exist for delivery of commodities within the period, and some forward markets exist for future delivery, but only relatively few. Hence, at any one time only a relatively small subset of all commodities can be exchanged. We then get a sequence of market systems, one in each period, and exchange activity takes place continually.

This picture of the economy raises a number of interesting questions. How will the outcomes on markets at one period be influenced by expectations about the outcomes in later periods? What will be the relationship, if any, between spot prices of commodities with the same physical attributes but different dates of delivery (e.g. the price of crude oil now and its price this time next year)? Can income (which we can take here to be the proceeds of sales of commodities, including of course labour services) be transferred between time periods and, if so, how? What are the consequences of the fact that the future cannot be known with certainty?

The analysis of the full implications of the view of the economy as a time sequence of market systems is complex and still incomplete. Our approach is to take it in three stages. We first analyse an 'atemporal economy', which could be thought of as an economy existing for just a single time period. We then extend the analysis to an 'intertemporal economy' by considering an economy which will exist over more than one period, but make the assumption of *complete certainty* – all relevant facts about the future are fully known at each point in time. We then take the final step of relaxing this certainty assumption and allowing incomplete information about data relating to the future. It is the analysis of this last kind of economy which is not yet complete. As long as we assume complete certainty, analysis of an intertemporal economy can be made formally identical to that of the atemporal economy, or, alternatively, identical to that of the kind of economy in which there is a complete system of spot and forward markets existing at any one time (see Chapter 15). At a more advanced level of analysis, it is usual to merge stages one and two, and analyse an economy which could be interpreted either atemporally or intertemporally. Indeed, on certain quite strong assumptions it is possible to do the same for the economy with uncertainty (see Chapter 20). However, in this book we shall take one stage at a time.

4. Economic agents

The basic units of analysis in microeconomics are the individual economic agents or decision-takers (hence the term *microeconomics*), who are usually classified either as

consumers or *firms*. A consumer is regarded as an individual who may own initially certain stocks of commodities, his 'initial endowment' (counted as part of his wealth), and who has to choose an amount of each commodity (which may of course be zero) to consume. This amount, in conjunction with his initial endowment, will determine the quantity of each commodity he will want to buy or sell on the relevant market. An alternative and less general formulation is to ignore the selling side of the consumer's activities, and assume his initial endowment takes the form of 'income', expressed in units of account or in terms of some numeraire. We then analyse simply his consumption (equals purchasing) decision, assuming also that he holds zero stocks of all the goods he might want to consume. This somewhat restrictive view of the consumer's activities is useful as a way of developing certain tools of analysis, but clearly can only be provisional, if we also want to say anything about the supply of commodities such as labour services.

A firm is also usually regarded as an individual decision-taker, undertaking the production of commodities by combining inputs in technological processes. These inputs will usually themselves be commodities, some of which the firm may own as part of its initial endowment, and some of which it may buy on the relevant markets. In certain cases, however, important inputs may not be commodities, e.g. sunshine in the production of wine. The crux of the distinction between consumers and firms is the nature of their economic activity: consumers buy and sell commodities in order to consume; firms buy inputs and produce commodities in order to sell.

In reality the counterparts of these theoretical abstractions are more complex. 'Consumer units' are usually groups of two or more people comprising a 'household' and decisions on purchases and sales may well be group decisions. Provided that the household acts in its decision-taking in a way which corresponds reasonably closely to certain principles of rationality and consistency, it is *enough for the purposes of our theory* to regard it as a single abstract decision-taker, 'the consumer'. If the organization of the household were shown to be such as to lead to significant departures from these principles of rationality and consistency, and to make our theories seriously misleading when used to explain and predict consumption decisions then we would have to reconstruct our theory of the consumer to take account of this.

In the case of the firm, the empirical counterpart of the theoretical entity may be thought even less like a single individual. Although many owner-controlled or *entrepreneurial* firms exist, economic activity is dominated by large corporations, with complex structures of organization and decision-taking. We can apply the same argument as before: it is a simplifying theoretical abstraction to ignore the organizational characteristics of firms for the purpose of our analysis of the general resource allocation process. This is defensible as long as the explanations and predictions we make about the decisions of firms in this process are not shown to be false by the evidence of firms' behaviour. Much more so than in the case of the household, however, there is a great deal of argument and some evidence to suggest that certain aspects of the organizational structure of firms *do* lead them to behave differently from the predictions of the theory of the firm as a single decision-taker. Accordingly in Chapters 6 and 13 we examine theories which take some account of the organizational characteristics of modern corporations.

The classification of the set of economic agents into consumers and firms reflects the basic distinction between the activities of production and consumption. We can quite easily choose a less rigid separation between types of economic agents. For example, if the

decision-taker controlling the firm is a person, the *entrepreneur*, then she is necessarily a consumer as well as a producer. We could then construct a theory which has the producer taking consumption decisions as well as production decisions. This leads to a view of an economy as consisting basically of consumers, at least some (and possibly all) of whom have access to production possibilities – they possess the knowledge, skills and initial endowment of commodities (including probably ‘capital goods’) which enable them to produce as well as exchange. Such an economy is quite amenable to analysis by the methods developed for the economy in which we preserve the distinction between consumers and producers. Indeed, if we make the assumption that inputs of ‘managerial services’ can be bought and sold on a market, there is no essential difference between the two economies.

An alternative way of blurring the distinction between consumers and producers is to regard the consumer as in fact a kind of producer. At the simplest level, we could view her as ‘producing’ labour services, using as inputs the commodities she buys, and with these labour services being supplied to firms which use them in conjunction with other commodities to produce commodities which are supplied to consumers ... and so on. The point of interest in this kind of economy is to study the conditions under which the economy can sustain or *reproduce* itself – will the flow of commodities as outputs be just sufficient to produce the labour services and other commodities necessary again to produce that same flow? We could also examine the way in which such an economy might grow over time by producing in each period more commodities than are required simply to reproduce themselves. A more sophisticated model of the consumer as a producer regards her as buying market goods and services, and combining them with her own time and effort, to ‘produce’ certain consumption services, which are the real objects of consumption. For example, a rail journey from *A* to *B* involves the purchase of a transportation service on the market, together with an input of the traveller’s time, to produce the consumption service of a trip from *A* to *B*. The method of analysis developed for production by firms is then used to analyse the consumer’s choices of market commodities when they are regarded as inputs into the production of consumption services. Such models have wide applications, for example in the analysis of markets for transport services of various kinds, where the duration of travelling time is an important aspect of the choice, and also to help us understand why, as real incomes increase, consumers appear to substitute time- and labour-saving commodities for others. Such models are useful whenever we want to bring to the forefront of the analysis the fact that time is a scarce resource.

5. Rationality

In whatever way we break down the distinction between consumers and producers in microeconomic models, two central elements remain. First is the adoption of the individual decision-taker as the basic unit of analysis. Second is the hypothesis that this decision-taker is *rational*. The concept of rationality is so pervasive that its meaning must be clearly expressed. We would say that rational decision-taking takes the following form:

- (a) The decision-taker sets out *all* the *feasible* alternatives, rejecting any which are not feasible;

- (b) He takes into account whatever information is readily available, or worth collecting, to assess the consequences of choosing each of the alternatives;
- (c) In the light of their consequences he ranks the alternatives in order of preference, where this ordering satisfies certain assumptions of completeness and consistency (discussed in Chapter 3 below);
- (d) He chooses the alternative highest in this ordering, i.e. he chooses the alternative with the consequences he prefers over all others available to him.

These ‘requirements of rationality’ seem to be quite consistent with the everyday sense in which rationality is used. People *can* behave irrationally in this sense: in taking a decision, they may ignore *known* feasible alternatives, they may allow themselves to be influenced by infeasible alternatives, they may ignore or not bother to collect information on the consequences of their decisions, they may contradict themselves in the ranking of the alternatives, and they may even choose an alternative whose consequences *they have already told us* they regard as less attractive than those of another alternative. That is to say, the assumption of rationality is an *hypothesis*, rather than a *tautology* – we can quite well conceive of its being false for a particular decision-taker.

However, it is not always as easy to conclude that a decision-taker is behaving irrationally as may be supposed. The important principle here is (b) above, relating to the use and acquisition of information. The collection of information, and the process of decision-taking itself, absorbs resources and therefore imposes costs. Given that all the information which could possibly be relevant to a decision is not readily and costlessly available, we may often observe behaviour which is rational on the basis of principles (a)–(d), but may be labelled irrational by a careless observer (or one determined to prove that *homo economicus* does not exist). For example, a consumer may habitually use the same supermarket rather than shopping around other supermarkets to find better bargains. This might appear to violate principle (a), but could be explained by the arguments that habit is essentially a way of economizing on time and effort, and that the consumer’s expectation of the gain he would make by shopping around does not seem to him to justify the cost and bother involved.

The danger in this kind of explanation is apparent in the example: with a little ingenuity, just about any kind of behaviour could be made to appear rational. This is a danger we have to avert, if the concept of rationality is not to become an empty tautology – we have to accept that people may at times be irrational. The general point can be put in the following way. We do observe in economic behaviour a tendency for the consistent pursuit of well-defined objectives. Consumers’ ‘habits’ have often been shown to be very easily changed by sufficiently large price changes. We also observe instances of apparent irrationality which may or may not be convincingly explained away. This therefore suggests that it is very difficult to test the hypothesis of rationality by actually observing individuals as they go through *the process* of decision-taking. For many purposes we may not even require that *every* individual act rationally, as long as in the aggregate enough people act with enough rationality to make our theories of the behaviour of these aggregates (e.g. all the buyers in a market) applicable. This suggests that the best *practical* test of the rationality hypothesis is by testing the further hypotheses which are derived from it,

It is important to distinguish between *forward* and *spot* markets. On a spot market, an agreement is made under which delivery of a commodity is completed within the current period; on a forward market, delivery will be made at some future period. (Some markets may do both, e.g. the market in leasehold accommodation, where what may be sold is a flow of housing services over possibly a very large number of years). We could envisage an economy in which at a given point in time there exists a market for every commodity, which means there is a complete system of spot and forward markets. In such an economy, contracts would be entered into for all future exchanges of commodities as well as for all current exchanges, and so market activity could cease entirely after the first period: the rest of the time would be spent simply fulfilling the contracts already concluded. Real economies, of course, do not possess such complete market systems. In any one period, markets exist for delivery of commodities within the period, and some forward markets exist for future delivery, but only relatively few. Hence, at any one time only a relatively small subset of all commodities can be exchanged. We then get a sequence of market systems, one in each period, and exchange activity takes place continually.

This picture of the economy raises a number of interesting questions. How will the outcomes on markets at one period be influenced by expectations about the outcomes in later periods? What will be the relationship, if any, between spot prices of commodities with the same physical attributes but different dates of delivery (e.g. the price of crude oil now and its price this time next year)? Can income (which we can take here to be the proceeds of sales of commodities, including of course labour services) be transferred between time periods and, if so, how? What are the consequences of the fact that the future cannot be known with certainty?

The analysis of the full implications of the view of the economy as a time sequence of market systems is complex and still incomplete. Our approach is to take it in three stages. We first analyse an 'atemporal economy', which could be thought of as an economy existing for just a single time period. We then extend the analysis to an 'intertemporal economy' by considering an economy which will exist over more than one period, but make the assumption of *complete certainty* – all relevant facts about the future are fully known at each point in time. We then take the final step of relaxing this certainty assumption and allowing incomplete information about data relating to the future. It is the analysis of this last kind of economy which is not yet complete. As long as we assume complete certainty, analysis of an intertemporal economy can be made formally identical to that of the atemporal economy, or, alternatively, identical to that of the kind of economy in which there is a complete system of spot and forward markets existing at any one time (see Chapter 15). At a more advanced level of analysis, it is usual to merge stages one and two, and analyse an economy which could be interpreted either atemporally or intertemporally. Indeed, on certain quite strong assumptions it is possible to do the same for the economy with uncertainty (see Chapter 20). However, in this book we shall take one stage at a time.

4. Economic agents

The basic units of analysis in microeconomics are the individual economic agents or decision-takers (hence the term *microeconomics*), who are usually classified either as

consumers or *firms*. A consumer is regarded as an individual who may own initially certain stocks of commodities, his 'initial endowment' (counted as part of his wealth), and who has to choose an amount of each commodity (which may of course be zero) to consume. This amount, in conjunction with his initial endowment, will determine the quantity of each commodity he will want to buy or sell on the relevant market. An alternative and less general formulation is to ignore the selling side of the consumer's activities, and assume his initial endowment takes the form of 'income', expressed in units of account or in terms of some numeraire. We then analyse simply his consumption (equals purchasing) decision, assuming also that he holds zero stocks of all the goods he might want to consume. This somewhat restrictive view of the consumer's activities is useful as a way of developing certain tools of analysis, but clearly can only be provisional, if we also want to say anything about the supply of commodities such as labour services.

A firm is also usually regarded as an individual decision-taker, undertaking the production of commodities by combining inputs in technological processes. These inputs will usually themselves be commodities, some of which the firm may own as part of its initial endowment, and some of which it may buy on the relevant markets. In certain cases, however, important inputs may not be commodities, e.g. sunshine in the production of wine. The crux of the distinction between consumers and firms is the nature of their economic activity: consumers buy and sell commodities in order to consume; firms buy inputs and produce commodities in order to sell.

In reality the counterparts of these theoretical abstractions are more complex. 'Consumer units' are usually groups of two or more people comprising a 'household' and decisions on purchases and sales may well be group decisions. Provided that the household acts in its decision-taking in a way which corresponds reasonably closely to certain principles of rationality and consistency, it is *enough for the purposes of our theory* to regard it as a single abstract decision-taker, 'the consumer'. If the organization of the household were shown to be such as to lead to significant departures from these principles of rationality and consistency, and to make our theories seriously misleading when used to explain and predict consumption decisions then we would have to reconstruct our theory of the consumer to take account of this.

In the case of the firm, the empirical counterpart of the theoretical entity may be thought even less like a single individual. Although many owner-controlled or *entrepreneurial* firms exist, economic activity is dominated by large corporations, with complex structures of organization and decision-taking. We can apply the same argument as before: it is a simplifying theoretical abstraction to ignore the organizational characteristics of firms for the purpose of our analysis of the general resource allocation process. This is defensible as long as the explanations and predictions we make about the decisions of firms in this process are not shown to be false by the evidence of firms' behaviour. Much more so than in the case of the household, however, there is a great deal of argument and some evidence to suggest that certain aspects of the organizational structure of firms *do* lead them to behave differently from the predictions of the theory of the firm as a single decision-taker. Accordingly in Chapters 6 and 13 we examine theories which take some account of the organizational characteristics of modern corporations.

The classification of the set of economic agents into consumers and firms reflects the basic distinction between the activities of production and consumption. We can quite easily choose a less rigid separation between types of economic agents. For example, if the

decision-taker controlling the firm is a person, the *entrepreneur*, then she is necessarily a consumer as well as a producer. We could then construct a theory which has the producer taking consumption decisions as well as production decisions. This leads to a view of an economy as consisting basically of consumers, at least some (and possibly all) of whom have access to production possibilities – they possess the knowledge, skills and initial endowment of commodities (including probably ‘capital goods’) which enable them to produce as well as exchange. Such an economy is quite amenable to analysis by the methods developed for the economy in which we preserve the distinction between consumers and producers. Indeed, if we make the assumption that inputs of ‘managerial services’ can be bought and sold on a market, there is no essential difference between the two economies.

An alternative way of blurring the distinction between consumers and producers is to regard the consumer as in fact a kind of producer. At the simplest level, we could view her as ‘producing’ labour services, using as inputs the commodities she buys, and with these labour services being supplied to firms which use them in conjunction with other commodities to produce commodities which are supplied to consumers ... and so on. The point of interest in this kind of economy is to study the conditions under which the economy can sustain or *reproduce* itself – will the flow of commodities as outputs be just sufficient to produce the labour services and other commodities necessary again to produce that same flow? We could also examine the way in which such an economy might grow over time by producing in each period more commodities than are required simply to reproduce themselves. A more sophisticated model of the consumer as a producer regards her as buying market goods and services, and combining them with her own time and effort, to ‘produce’ certain consumption services, which are the real objects of consumption. For example, a rail journey from *A* to *B* involves the purchase of a transportation service on the market, together with an input of the traveller’s time, to produce the consumption service of a trip from *A* to *B*. The method of analysis developed for production by firms is then used to analyse the consumer’s choices of market commodities when they are regarded as inputs into the production of consumption services. Such models have wide applications, for example in the analysis of markets for transport services of various kinds, where the duration of travelling time is an important aspect of the choice, and also to help us understand why, as real incomes increase, consumers appear to substitute time- and labour-saving commodities for others. Such models are useful whenever we want to bring to the forefront of the analysis the fact that time is a scarce resource.

5. Rationality

In whatever way we break down the distinction between consumers and producers in microeconomic models, two central elements remain. First is the adoption of the individual decision-taker as the basic unit of analysis. Second is the hypothesis that this decision-taker is *rational*. The concept of rationality is so pervasive that its meaning must be clearly expressed. We would say that rational decision-taking takes the following form:

- (a) The decision-taker sets out *all* the *feasible* alternatives, rejecting any which are not feasible;

- (b) He takes into account whatever information is readily available, or worth collecting, to assess the consequences of choosing each of the alternatives;
- (c) In the light of their consequences he ranks the alternatives in order of preference, where this ordering satisfies certain assumptions of completeness and consistency (discussed in Chapter 3 below);
- (d) He chooses the alternative highest in this ordering, i.e. he chooses the alternative with the consequences he prefers over all others available to him.

These ‘requirements of rationality’ seem to be quite consistent with the everyday sense in which rationality is used. People *can* behave irrationally in this sense: in taking a decision, they may ignore *known* feasible alternatives, they may allow themselves to be influenced by infeasible alternatives, they may ignore or not bother to collect information on the consequences of their decisions, they may contradict themselves in the ranking of the alternatives, and they may even choose an alternative whose consequences *they have already told us* they regard as less attractive than those of another alternative. That is to say, the assumption of rationality is an *hypothesis*, rather than a *tautology* – we can quite well conceive of its being false for a particular decision-taker.

However, it is not always as easy to conclude that a decision-taker is behaving irrationally as may be supposed. The important principle here is (b) above, relating to the use and acquisition of information. The collection of information, and the process of decision-taking itself, absorbs resources and therefore imposes costs. Given that all the information which could possibly be relevant to a decision is not readily and costlessly available, we may often observe behaviour which is rational on the basis of principles (a)–(d), but may be labelled irrational by a careless observer (or one determined to prove that *homo economicus* does not exist). For example, a consumer may habitually use the same supermarket rather than shopping around other supermarkets to find better bargains. This might appear to violate principle (a), but could be explained by the arguments that habit is essentially a way of economizing on time and effort, and that the consumer’s expectation of the gain he would make by shopping around does not seem to him to justify the cost and bother involved.

The danger in this kind of explanation is apparent in the example: with a little ingenuity, just about any kind of behaviour could be made to appear rational. This is a danger we have to avert, if the concept of rationality is not to become an empty tautology – we have to accept that people may at times be irrational. The general point can be put in the following way. We do observe in economic behaviour a tendency for the consistent pursuit of well-defined objectives. Consumers’ ‘habits’ have often been shown to be very easily changed by sufficiently large price changes. We also observe instances of apparent irrationality which may or may not be convincingly explained away. This therefore suggests that it is very difficult to test the hypothesis of rationality by actually observing individuals as they go through *the process* of decision-taking. For many purposes we may not even require that *every* individual act rationally, as long as in the aggregate enough people act with enough rationality to make our theories of the behaviour of these aggregates (e.g. all the buyers in a market) applicable. This suggests that the best *practical* test of the rationality hypothesis is by testing the further hypotheses which are derived from it,

especially those further hypotheses which could not be derived from a postulate of 'irrationality' (somehow specified).

To summarize the discussion of this chapter so far, the basic elements of microeconomics are: *rational* individual decision-takers, usually classified as consumers and firms; commodities; markets; prices.

6. Method of analysis

The core of microeconomic theory follows through a systematic line of development. We begin with models of the individual decision-takers, a 'typical' or representative consumer and a 'typical' or representative firm. The assumption of rationality implies that these models take the form of *optimization problems*: the decision-taker is assumed to seek the *best* alternative out of the feasible set of alternatives open to him. By specifying fairly closely the nature of these optimization problems and then solving them, we are able to attribute certain characteristics and properties to the decision-taker's choices. Moreover, by examining the way in which the optimal choices may vary with changes in underlying parameters of the decision problem (especially prices), we can trace out *behaviour relationships* such as demand and supply curves.

A major purpose of the models of individual decisions is to allow us to place certain restrictions on these behaviour relationships, or at least to clarify the assumptions under which particular restrictions (e.g. that demand curves have negative slopes) can be placed.

The next step in the development of the theory is to aggregate the individual behaviour relationships over groups of economic agents – usually the set of buyers in a market on the one hand and the set of sellers in a market on the other, where these can be separately identified. This latter qualification is necessary because, as we shall see, in some market models an individual may be a buyer at some prices and a seller at others, so that no hard and fast distinction can be drawn between buyers and sellers, and aggregation takes place over *all* economic agents. These aggregated relationships then form the basis for an analysis of the operation of a single market taken in isolation, and also of systems of several interrelated markets. At the most general, we consider the system of markets for the economy as a whole, and analyse the way in which a resource allocation is determined by the simultaneous interaction of this market system.

The method of analysis is the same throughout, and can be described as the *equilibrium methodology*. The equilibrium of a system is defined as a situation in which the forces determining the state of that system are in balance, so that there is no tendency for the variables of the system to change. (*Note*: Strictly speaking, this is the method of *static* equilibrium analysis. We could allow variables and parameters to vary with time, and look for *equilibrium time-paths*, in a dynamic analysis. Since in this book we use static methods throughout, we do not need the qualification.) An equilibrium of a system of economic agents (which may be a single market or a whole economy) will exist when two conditions are satisfied:

- (a) individual decision-makers have no wish to change their planned decisions or reactions;
- (b) the plans of decision-makers are consistent or compatible and hence can be realized.

The significance of the equilibrium concept is that it provides us with a *solution principle*. Once we have defined the forces operating within a given economic system, for example a model of a single market, we naturally ask the question: what will the outcome of the interaction of those forces be? The answer is provided by the concept of equilibrium: we find the characteristics of the equilibrium state of the system, and take this as the outcome we seek. But if we want to use the equilibrium state as a prediction of the outcome of the workings of the system, we first have to answer a number of fundamental questions:

- (a) *Existence*. Does the system in fact *possess* an equilibrium state, i.e. given the forces operating within the system, is there in principle a state in which they would be in balance, or is it the case that no such state of balance is possible? Clearly, if a system does not possess an equilibrium, we cannot describe its outcome as an equilibrium state.
- (b) *Stability*. Suppose that an equilibrium state does exist. Then, given that the system may not initially be in this state, would it tend to converge to it? If it does, then we call the system *stable*. Clearly, the equilibrium state loses much of its interest if the system is not in this sense stable, since it is unlikely ever actually to be attained.
- (c) *Uniqueness*. A system may possess more than one equilibrium state, and the different possible equilibria may have different properties and implications, and so it is of interest to know for a given system whether there is only one possible equilibrium state which needs to be described.

These questions of the existence, stability and uniqueness of an equilibrium state are necessarily raised by use of an equilibrium methodology, and so we shall find that we shall be considering them in a number of contexts throughout this book.

These introductory remarks have been concerned with describing the basic concepts of microeconomics, the overall structure of the theory and its method of analysis. Full understanding of these can only be achieved, of course, by a thorough study of the theory itself. Before embarking upon this, we conclude this introductory chapter with some comments on the view of the economic and social system which is implicit in modern microeconomic analysis.

B. The economic and social framework

We do not say very much in the rest of this book about the institutional, political and legal framework within which our economic analysis is set. Much of microeconomic theory implicitly assumes a certain kind of framework and is concerned with examining the major economic forces which operate within it. Despite this the theory can offer great insights into a variety of institutional frameworks. Some fundamental economic issues exist whatever the institutional form and one means of comparing alternative systems is in terms of the way in which these problems manifest themselves and are dealt with.

Three facts of economic life appear in all types of society. The first is *relative scarcity* of resources: however abundant in absolute terms are the resources possessed by a society the individuals in the society want to consume more goods and services than can be

produced from those resources. Second, there are gains from *specialization*: the output of goods and services will be greater if individuals specialize in different aspects of the production process and each does not attempt to produce all the commodities they consume. Third, *information is decentralized*: no single individual initially knows all the economically relevant information. This information includes both the characteristics of individuals such as their preferences and their endowments of resources (widely defined to include their skills, productivity, the quality of goods they have to sell) and the actions they take (for example how hard or carefully they work). Given these facts, every society is faced with the problems of organizing exchange and coordinating the separate decisions of the large numbers of consumers and producers.

The decisions taken by individuals in an economy are constrained both by *technology* and by the set of *property rights*. Technological constraints arise from the fundamental physical laws which determine what outputs of goods and services can be produced from given sets of resources. Property rights are the rules (whether formal and legal or informal custom) which specify what individuals are allowed to do with resources and the outputs of those resources. Property rights define which of the technologically feasible economic decisions individuals are *permitted* to make.

The institutional frameworks of economies can be classified by the sets of property rights with which they are associated. The microeconomic theory in this book was originally developed to examine how the basic economic problems are solved in a *decentralized private ownership economy*. In such an economy the set of property rights vests ownership of resources and commodities in individuals. All resources are owned by specified individuals who have the right to use them for a wide variety of purposes and can sell that right to other individuals. Decisions are decentralized in the sense that there is no agency or individual in the economy with the right to tell any individual what she must do with the resources she owns. The state's role in such economic models is minimal: it is tacitly assumed to enforce and define the set of private ownership rights and to provide the institutions this requires: a police force and civil and criminal courts.

Microeconomic analysis, however, has relevance beyond its application to such an economy. The concepts and techniques developed in this book can be used to examine the behaviour of individuals in economies with a wider role for the state and with other institutional frameworks. Economies with different institutional frameworks impose different constraints on decisions because of the differences in the sets of property rights, but the basic microeconomic methodology is unchanged. Individuals in such economies can still be modelled as rational agents optimizing subject to constraints and thus we can make predictions about how their behaviour responds to changes in their environment. We can still define an equilibrium in such economies as a situation in which individuals make optimal decisions which are mutually consistent and thus can be implemented. The equilibrium we investigate may not look like the equilibrium of the simple decentralized private ownership economy, but we can still investigate the circumstances under which it will exist, be stable and be unique, compare the equilibria which arise as conditions change and make welfare judgements about the resulting allocations. Thus in an economy with prices fixed by a central authority, equilibrium may be compatible with consumers spending considerable lengths of time waiting in line in order to acquire commodities. This situation could not be an equilibrium in an unregulated private ownership economy but this does not mean that we cannot use microeconomic theory to examine it.

Microeconomic theory has been used to examine the allocation of resources in feudal economies, slave economies, centrally planned economies, co-operative economies and in mixed economies where private ownership is combined with a large state sector and extensive regulation of individual decision-making. In short, the microeconomic methodology we set out can be used to analyse economic decisions in a wide range of institutional frameworks and to examine the consequences of changes in those frameworks.

References and further reading

There are critical discussions of rationality as a basis for the analysis of decisions in:

- J. Elster, *Solomonic Judgements: Studies in the Limits of Rationality*, Cambridge University Press, Cambridge, 1989.
- H. A. Simon, 'Theories of bounded rationality', in G. B. McGuire and R. Radner (eds), *Decision and Organisation*, North Holland, London, 1972.
- K. S. Cook and M. Levi (eds) *The Limits of Rationality*, Chicago University Press Chicago, 1990.

The extension of the equilibrium concept to dynamic economies under uncertainty is discussed in:

- F. H. Hahn, 'On the notion of equilibrium in economics' in F. H. Hahn, *Equilibrium and Macroeconomics*, Basil Blackwell, Oxford, 1984.

Alternative 'Austrian' views of rational behaviour and equilibrium are presented in:

- L. L. Lachman, *The Market as an Economic Process*, Basil Blackwell, Oxford, 1986.
- B. J. Loasby, *Choice, Complexity and Ignorance*, Cambridge University Press Cambridge, 1976.

A general survey of the methodology of economics is provided in:

- M. Blaug, *The Methodology of Economics*, Cambridge University Press, Cambridge, 1980.

CHAPTER 2

Optimization

As the previous chapter has suggested, the idea of rationality in economics implies, among other things, that a decision-taker tries to find the best alternative out of those available to him. In other words he tries to *optimize*. For this reason the idea of optimization is fundamental to microeconomics and optimization problems occur repeatedly throughout this book. Although the particular economic context of the problems may vary – a consumer's choice of consumption bundle, a firm's production decision, a planner's choice of resource allocation in the economy, for example – they tend to have a common basic structure. It is therefore efficient and illuminating to invest some time in considering optimization problems in general, abstracting from any particular context.

An examination of the general theory of optimization problems puts the standard models of economics in context. It makes us aware of the assumptions which have been quietly slipped into the analysis to ensure that the results have certain properties, and enables us to reformulate the answers when these implicit assumptions have been removed. At the very least it greatly increases our awareness of the nature and meaning of the economic models themselves.

A. The structure of an optimization problem

We have defined optimization simply to mean the act of choosing the 'best' alternative out of whatever alternatives are available. It is a description of how decisions (choices among alternatives) are or should be taken. We now go beyond this simple idea and examine in some detail the questions we can ask about optimization problems and the concepts which have been developed to answer them.

All optimization problems consist of three elements:

1. Choice variables

These are the variables whose optimal values have to be determined. For example:

- (a) A firm wants to know at what level to set output in order to achieve maximum profit. Output is the choice variable.

- (b) A firm wants to know what amounts of labour, machine time and raw materials to use so as to produce a given output level at minimum cost. Choice variables are labour, machine time, raw materials.
- (c) A consumer wants to buy that bundle of commodities which he can afford and which makes him feel best off. Here the choice variables are quantities of commodities.

In economics the amount of any choice variable is almost always assumed to be measurable as a real number. There may be any finite number of choice variables in a particular problem.

2. The objective function

This gives a mathematical specification of the relationship between the choice variables on the one hand and some variable whose value we wish to *maximize* or *minimize* on the other. Thus, in the three examples just discussed, the objective functions would relate:

- (a) profit to the level of output;
- (b) cost to the amounts of labour, machine time and raw materials;
- (c) an index of the consumer's satisfaction to the quantities of the commodities he may buy.

In (a) and (c) the functions are to be maximized, and in (b) minimized, with respect to the relevant choice variables.

The reader has, we hope, sensed a difference between the third of these objective functions and the first two. Profit and cost are money magnitudes, the measurement of which seems to present no problems. What, however, do we mean by 'an index of satisfaction'? Satisfaction is an internal subjective thing, and it is not immediately obvious that they can be represented by a numerical index. In the next chapter we shall consider at some length a set of assumptions which, if they hold, allow us to conclude that in the consumer's problem, we can define a numerical objective function which he is taken to maximize.

Ultimately, all objective functions, even those which appear at first sight to involve readily measurable magnitudes like profit and cost, are numerical representations of preference orderings. An entrepreneur aims to maximize profit because he *prefers* more profit to less, in terms of the uses to which it may be put. He seeks to minimize cost because he *prefers* to pay out less rather than more, given the other things he may spend money on. These two objective functions in fact correspond to a rather special case, in which the decision-taker is interested in only a single measurable outcome of the decision (profit or cost), and his preference ordering over values of this is of a particularly simple form – he always prefers more to less, or less to more. In such a special case, we can ignore the order of preference and operate in terms of the single outcome directly. If, on the other hand, more than one outcome of the decision were to be relevant (e.g. both profit and the amount of effort put in), then we cannot ignore the more fundamental preference ordering. One alternative may involve more of one variable and less of the other than a second alternative, and so some kind of relative evaluation is inescapable.

For the rest of this chapter, we shall simply take it that we wish to maximize some magnitude which is a real number, and which is a given function of the choice variables in the problem. The foregoing discussion, however, gives us a useful interpretation of the objective function: we can view it as placing alternative values of the choice variables in order of preference so that we can find those preferred.

3. The feasible set

So far we have talked loosely about 'the available alternatives'. An essential part of any optimization problem is a specification of exactly what alternatives are available to the decision-taker. The available set of alternatives is called the 'feasible set'. Since the alternatives are usually regarded as points, feasible sets are usually point sets.

There are three ways in which the feasible set may be specified:

1. By direct enumeration, i.e. by a statement which says: the alternatives are A, B, C, \dots . Clearly, if the choice set contains one alternative, the optimization problem is trivial and if none, it is insoluble.
2. By one or more inequalities which *directly* define a set of alternative values of the choice variable(s).
3. By one or more *functions* or equations which define a set of alternative values.

Examples of the last two of these can be found in the problems discussed earlier. Thus, in problem (a), we would rule out negative outputs, but would expect any positive outputs to be possible. Hence we would say that output must be greater than or equal to zero, i.e. $y \geq 0$, where y is output. The feasible set is here directly defined by a *weak inequality*.

In the second problem, we are told that only those combinations of inputs which yield the desired output level can be considered. In this case the feasible set is generally defined by a function. Thus, suppose we have the *production function*: $y = f(L, M, R)$ where y is output, L, M, R , are labour, machine time and raw materials respectively. Now let y^0 be the required output level. Then the equation:

$$y^0 = f(L, M, R) \quad [A.1]$$

defines a set of values of L, M, R , which are feasible. In addition, note that it may be possible for equation [A.1] to be satisfied by negative values of one or more of the choice variables, implying that such negative values may be chosen. We would not regard negative values of these variables as making sense, however, and so we wish to exclude them. Thus, we would add the direct constraints:

$$L \geq 0 \quad R \geq 0 \quad M \geq 0 \quad [A.2]$$

which, in conjunction with [A.1] define the feasible set.

In the example of the consumer's problem we can say first of all that it is impossible to consume negative amounts of goods. Then, if we let x_1, \dots, x_n represent the quantities of the goods which the consumer could buy, we have immediately the n inequalities:

$$x_1 \geq 0, \quad x_2 \geq 0, \quad x_3 \geq 0, \quad \dots, x_n \geq 0 \quad [A.3]$$

But there is another limitation on the feasible set implicit in the problem. Each good has a price. Let these prices be p_1, p_2, \dots, p_n , for x_1, \dots, x_n respectively. Then the consumer's total expenditure for some set of quantities of the goods will be:

$$p_1 x_1 + p_2 x_2 + \dots + p_n x_n = \sum_{i=1}^n p_i x_i \quad [A.4]$$

The consumer will have a given income, M , which consumption expenditure cannot exceed. Hence, in the problem, the feasible set is given by

$$p_1 x_1 + p_2 x_2 + \dots + p_n x_n \leq M \quad [A.5]$$

together with the inequalities in [A.4].

Thus, in each of the problems used as examples, as in any optimization problem, it is necessary to define in an exact way the available set of alternatives. The functions and inequalities written out above which limit or constrain the alternatives which can be considered in defining the feasible set are known as 'constraints'. If no constraints exist in a problem then the feasible set consists of the entire n -dimensional space of real vectors (where n is the number of choice variables) and the problem is called 'unconstrained'; the existence of constraints confines the feasible set to a subset of the whole space.

To summarize: an optimization problem consists of choice variables, an objective function, and a feasible set. The problem is to choose the preferred alternative in the feasible set, and our theory in general allows us to represent this as the problem of finding the maximum or minimum of the objective function with respect to the choice variables, subject to constraints. For this reason, optimization is taken to be synonymous with constrained maximization or minimization.

Exercise 2A

1. Describe the choice variables, objective functions and feasible sets in the following optimization problem:

- (a) You may go from college to home on foot, by bus, or by train. You know with certainty how long it takes by each mode, and what it will cost. The problem is to choose the 'best' way of going home.
- (b) You want to go on a diet which will involve as few calories as possible, subject to a certain minimum, and which will ensure that you consume at least minimum amounts of vitamins. You also cannot exceed your weekly food expenditure budget. You know the price, calorie count, and vitamin content of one unit of every foodstuff. There also exists a calorie-free but expensive all-vitamin tablet. How do you choose the 'best' diet?
- (c) A consumer may shop at market A, which is very close to his home, or take a bus ride and shop at market B, where prices are relatively lower. The problem is to choose whether to shop at market A, market B, or both.

2. Draw graphs of the feasible sets defined by the following constraints:

$$(a) \quad x_2 + 2x_1 \leq 4 \quad (b) \quad x_2 + 3x_1 < 6$$

$$x_1 \geq 0 \quad x_2 \geq 0$$

$$(c) \begin{aligned} x_2 + 2x_1 &\leq 4 \\ x_2 + 4x_1 &\leq 6 \\ x_1 &\geq 0 \quad x_2 &\geq 0 \end{aligned}$$

$$(e) \begin{aligned} x_2 + 2x_1 &= 4 \\ x_2 + 3x_1 &= 7 \\ x_1 &\geq 0 \quad x_2 &\geq 0 \end{aligned}$$

$$(g) \quad x_2 + x_1^2 = 4$$

$$(i) \begin{aligned} x_2 + x_1^2 &\leq 4 \\ x_2 + 3x_1 &\leq 6 \\ x_1 &\geq 0 \quad x_2 &\geq 0 \end{aligned}$$

$$(d) \begin{aligned} x_2 + 2x_1 &= 4 \\ x_2 + 3x_1 &= 7 \end{aligned}$$

$$(f) \begin{aligned} x_2 - 2x_1^2 &\leq 0 \\ x_2 &\geq 0 \quad x_1 &\geq 0 \end{aligned}$$

$$(h) \begin{aligned} x_2 + x_1^2 &= 4 \\ x_1 &\geq 0 \quad x_2 &\geq 0 \end{aligned}$$

$$(j) \begin{aligned} x_2 + x_1^2 &= 4 \\ x_2 + 3x_1 &= 6 \end{aligned}$$

B. Solutions: questions and concepts

A *solution* to an optimization problem is that vector of values of the choice variables which is in the feasible set and which yields a maximum or minimum of the objective function over the feasible set. It is useful here to introduce some notation. We present the objective function as:

$$f(x_1, x_2, \dots, x_n) = f(x) \quad [\text{B.1}]$$

where x is the n -component vector of choice variables. For convenience, we assume that the problem is always to *maximize* f .[†] We denote the feasible set of x vectors by S . Then a solution to the problem is a vector of choice variables, x^* , having the property:

$$f(x^*) \geq f(x) \quad \text{all } x \in S, \quad [\text{B.2}]$$

which is another way of saying that x^* maximizes f over the set S . By definition of the problem, we are interested in finding such a vector x^* .

There are certain important general questions we can ask about the solution to any optimization problem:

1. Existence

How can we be sure, in advance of trying to solve a particular problem, that a solution to it actually exists? After all, we have no grounds for supposing that every problem *must* have a solution. In economic theory, we spend a great deal of time analysing solutions to optimization problems. We therefore have to take care that theories provide for their existence, otherwise the analysis is internally inconsistent.

2. Local and global solutions

A *global* solution is one which satisfies the condition [B.2]; at that point, the objective function takes on a value which is not exceeded at any other point within the feasible set.

[†] Superior numerals are references to Notes at the end of the chapter.

It is therefore the solution we seek. A local solution, on the other hand, satisfies the condition:

$$f(x^{**}) \geq f(x) \quad \text{all } x \in N^{**} \in S \quad [\text{B.3}]$$

where N^{**} is a set of points in a *neighbourhood* of x^{**} . Figure 2.1 illustrates the difference. We assume only one choice variable, so that the vector becomes the scalar x . The feasible set is defined only by the direct inequalities: $x \geq 0$, $x \leq x^0$. The objective function $f(x)$ has two peaks, one at x^* and the other at x^{**} . Neighbourhoods of these points are shown as N^* and N^{**} respectively. Clearly, both points satisfy [B.3], but only x^* satisfies [B.2]. Thus, x^* is a global maximum, while x^{**} is not.

The difficulty is that *all the methods we have for finding solutions to optimization problems locate only local maxima*. We are therefore interested in the question: under what conditions will every local maximum we locate also be a global maximum? From Fig. 2.1 we see that this must have something to do with the shape of the objective function (suppose, for example, that the function had only a single peak); we shall explore this question in more detail below.

3. Uniqueness

It is conceivable that more than one global maximum may exist (e.g. suppose the first peak in Fig. 2.1 is as high as the second, or that the function has a horizontal segment over the set N^* of values of x). Economists have tended to assume unique solutions, and so it is of interest to consider conditions under which this is the case.

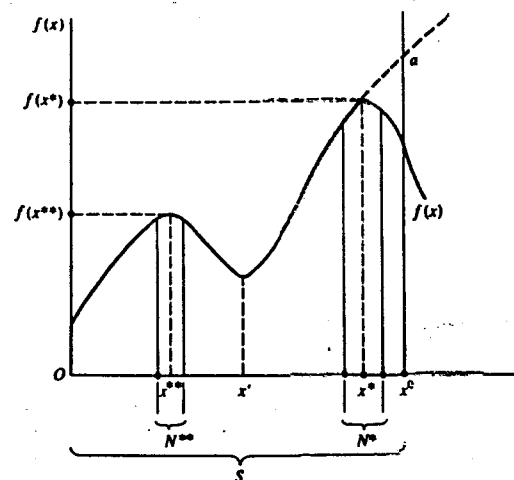


Fig. 2.1

4. Interior solutions

We take the distinction between an interior and a boundary point of a set as understood for the moment. Thus, in Fig. 2.1, the points $x = 0$ and $x = x^0$ are boundary points, while all other points in the set are interior points. Then an *interior solution* is an interior point which satisfies condition [B.2], while a *boundary solution* is a boundary point which satisfies that condition. In Fig. 2.1, x^* is an interior solution. If, however, the function f took the shape indicated by the dotted line in the diagram, then there would be a boundary solution at x^0 . The importance of the distinction relates to the question of the consequences of a change in a constraint (which in general changes the location of a boundary of the feasible set – refer back to Question 2, Exercise 2A). For a small change in a constraint, an interior solution is unlikely to be affected – the optimal point is unchanged. On the other hand, a boundary solution may well be affected – for example, in the case in which x^0 in Fig. 2.1 is optimal, a shift of the boundary would change the solution. Much of microeconomics is concerned with predictions of behaviour derived from an analysis of the change in optimal solutions following from a change in a constraint (for example see the analysis of consumer demand in the next chapter). It is therefore of importance to know whether a solution will be at a boundary or interior point.

We can also frame this question in terms of *binding* and *non-binding* constraints. A constraint is binding if there is a boundary solution which lies on the part of the boundary defined by that constraint. A constraint is non-binding therefore if there is an interior solution, or if the boundary solution lies on a part of the boundary defined by another constraint. For example, in Fig. 2.1, when the solution is at x^* , both constraints $x \geq 0$ and $x \leq x^0$ are non-binding while in the case in which the solution is at x^0 , the former constraint only is non-binding. We can always find a sufficiently small change in a non-binding constraint to leave the solution unaffected.

5. Location

Given that a solution exists, we would like to find it. In solving a practical problem we would obtain numerical values for the objective function and constraints and then try to devise computational procedures which will find a solution as quickly and cheaply as possible. In the theoretical context, however, we work only with general functions, usually specifying little more than the signs of their first and second derivatives. As a result, the problem is not actually to *compute* solutions but rather to *describe* their essential general characteristics in the analytically most useful way, in terms of certain conditions they have to satisfy.

To illustrate: take the case of unconstrained maximization, and suppose that in Fig. 2.1 no constraints on the feasible values of x exist. Then, at the value x^* , the derivative $f'(x^*) = 0$, and we can show that this must be true at all local maxima. However, it is also true that $f'(x') = 0$, but the function takes on a local minimum at x' . Hence, the condition that the first derivative be zero is satisfied at all local maxima but at other points also and so is necessary but not sufficient. We further note that as x increases through x^* , the derivative $f'(x^*)$ passes from positive to negative values, i.e. is decreasing, and that this is only true at a local maximum. It follows that a sufficient condition for a local maximum is that $f'(x^*) = 0$ and $f''(x^*) < 0$, since these are only satisfied at local maxima.²

Note that the conditions are defined only for *local* maxima, since, as the diagram makes clear, they cannot discriminate between points x^* and x^{**} . To describe an optimal point in terms of necessary and sufficient conditions is to 'locate' that point in terms of its general characteristics rather than its numerical value. Rather surprisingly perhaps, we are often able to say a very great deal on the basis of such a general description.

In discussing these questions concerning solutions to optimization problems, we make use of certain very general properties of objective functions and feasible sets. We shall first set out these properties, and then proceed to answer the questions.

1. Continuity of the objective function

A function $y = f(x)$ is continuous if there are no breaks in its graph, or crudely, if it can be drawn without taking the pen from the paper. In Fig. 2.2 the functions drawn in (b) and (c) are not continuous, while that in (a) is continuous. In (b) $f(x)$ becomes arbitrarily large at x^0 (tends to infinity) and in part (c) $f(x)$ jumps from y^1 to y^2 at x^0 . When there is more than one variable in the objective function the intuitive idea of continuity is still valid: there should be no jumps or breaks in the graph of the function.

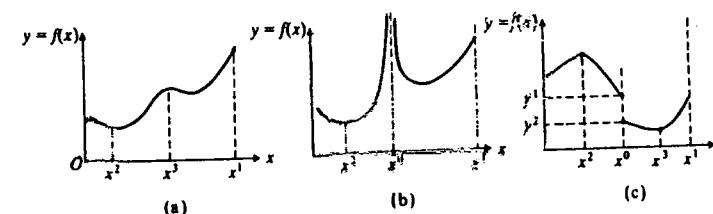


Fig. 2.2

2. Concavity of the objective function

In Fig. 2.3 we show graphs of four kinds of function. A function with the curvature shown in (a) of the figure would generally be called *concave*, that in (c) *convex*, while that in (b) is of course linear and that in (d) neither convex nor concave. If we wanted to find a non-geometrical way of defining the concave function one possibility would be in terms of the function's second derivative $f''(x)$. We note that as we draw successive tangents to the curve in (a) of the figure, at increasing values of x , these tangents have flatter and flatter positive slopes and then steeper and steeper negative ones, implying that the first derivative of the function, $f'(x)$, is decreasing. Thus we could express concavity by the condition $f''(x) < 0$. By a similar argument, convexity could be expressed by the condition $f''(x) > 0$. There is, however, a drawback to this. A function may not be differentiable at some point(s), because it possesses a kink there, and so the definition cannot be applied (draw an example).

To overcome this drawback we define concave and convex functions in terms of an obvious general property they possess. Note that in (a) of the figure, if we take any two

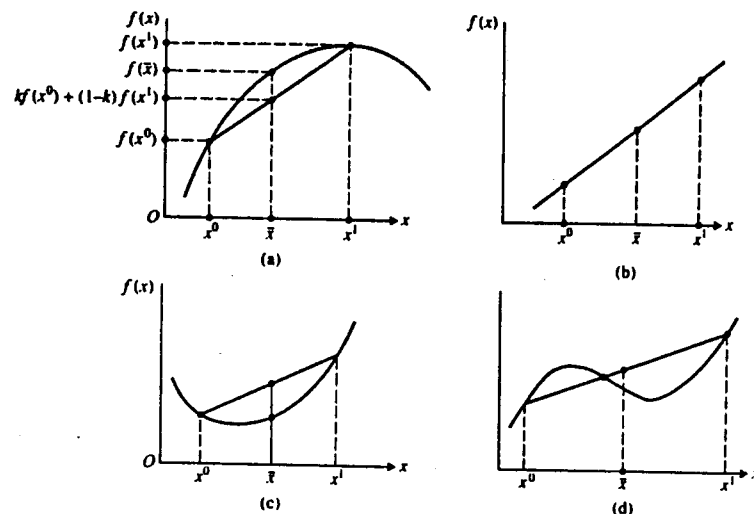


Fig. 2.3

points such as x^0 and x^1 , and join the corresponding function values $f(x^0)$ and $f(x^1)$ by a straight line (a chord to the function), then the graph of the function between these values lies everywhere above the line. For the convex function, the graph lies entirely below the line. In the linear case of course the graph of the function coincides with the line while in (d) of the figure it moves from above to below it.

In order to express this geometric idea algebraically, we note two things:

1. We can express any x -value lying between x^0 and x^1 as the weighted sum $\bar{x} = kx^0 + (1-k)x^1$, where k is between 0 and 1. For example $k = \frac{1}{2}$ gives us the x -value lying mid-way between x^0 and x^1 . The value \bar{x} is called the *convex combination* of the points x^0 and x^1 .
2. If we take the weighted sum of the two function values $f(x^0)$ and $f(x^1)$ using the same value of k , which we denote by $\bar{f} = kf(x^0) + (1-k)f(x^1)$, then this value is found as the co-ordinate of the point on the chord directly above \bar{x} . For example, for $k = \frac{1}{2}$, \bar{f} would lie at a point on the chord directly above $\bar{x} = \frac{1}{2}x^0 + \frac{1}{2}x^1$.⁽³⁾

For a concave function a point on the curve at any \bar{x} between x^0 and x^1 lies above the chord joining $f(x^0)$ and $f(x^1)$. It therefore follows from 2 that in this case $f(\bar{x}) > \bar{f}$, for all x lying between x^0 and x^1 . For a convex function we have $f(\bar{x}) < \bar{f}$ at each \bar{x} between x^0 and x^1 . When the function is linear $f(\bar{x}) = \bar{f}$ as Fig. 2.3(b) shows.

It turns out that some important propositions which are true when objective functions are shaped as in Fig. 2.3(a) are also true when they are linear or have linear segments. It is useful therefore to define *concave functions* as functions which have the property

$$f(\bar{x}) \geq \bar{f} \quad [\text{B.4}]$$

where $\bar{x} = kx^0 + (1-k)x^1$, and $\bar{f} = kf(x^0) + (1-k)f(x^1)$, $0 \leq k \leq 1$, so that linear functions, or functions with linear segments, may also be regarded as concave. A function which satisfies [B.4] as a strict inequality is then called *strictly concave*, and so this is the term which would be applied to the function shown in Fig. 2.3(a). Likewise a *convex function* satisfies:

$$f(\bar{x}) \leq \bar{f} \quad [\text{B.5}]$$

with a *strictly convex* function satisfying this as a strict inequality. Note that [B.4] and [B.5] taken together imply that a linear function is *both* convex and concave, though neither strictly convex nor strictly concave.

3. Quasi-concave functions

Given a function $y = f(x_1, x_2, \dots, x_n) = f(x)$, we can choose some number c , and let:

$$f(x) = c \quad [\text{B.6}]$$

This defines a set of values of the vector x having the property that they all yield the same value c of the function; in other words, they are solutions to the equation in [B.6]. Now a contour on a map is a set of points of equal height. Hence a useful terminology would be to say that [B.6] defines a *contour of the function* $f(x)$ and to call the set of x -values which satisfy [B.6] a *contour set*, since they correspond to equal values of the function. It turns out that for many purposes we are mainly interested in the *properties of contours of the objective function* and so we now want to consider the most important of these.

It is most convenient to take the two-variable case, so we assume that $x = (x_1, x_2)$. Thus, we have the contour:

$$f(x_1, x_2) = c \quad [\text{B.7}]$$

The advantage of taking the two-variable case is that we can graph the contour set satisfying [B.7] in two dimensions. A wide range of shapes is of course possible, and in Fig. 2.4 we present two examples. It is understood that all the points on a contour line c are vectors x which satisfy an equation for a contour such as [B.7], and so belong to a given contour set. The diagram illustrates one important property of a contour of a function, namely that of continuity. Just as before, continuity can be thought of intuitively as the absence of breaks, gaps, or jumps in the graph. Continuity of a function and of its contours are closely related: continuity of the function implies continuity of its contours.

To explore further the properties of contours, let us go beyond continuity and assume that the function $f(x)$ is differentiable. Then differentiating [B.7] totally:

$$\begin{aligned} df &= f_1 dx_1 + f_2 dx_2 \\ &= 0 \quad \text{since } c \text{ is a constant.} \end{aligned} \quad [\text{B.8}]$$

Note that the differentials dx_1 and dx_2 must be such as to keep the value of the function unchanged at c . Rearranging [B.8] gives:

$$\frac{dx_2}{dx_1} = -f_1/f_2 \quad (f_2 \neq 0) \quad [\text{B.9}]$$

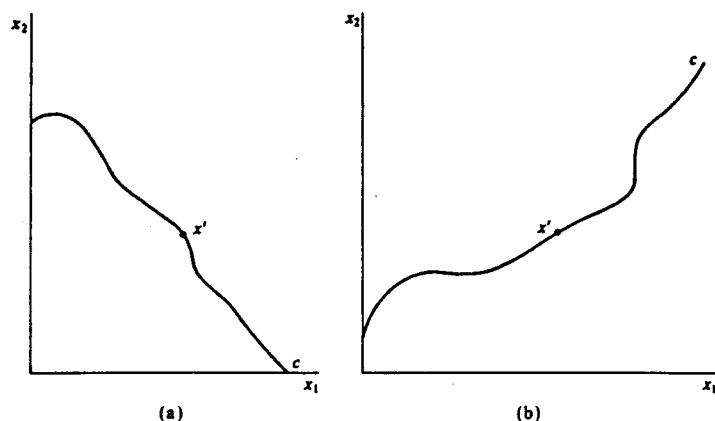


Fig. 2.4

The ratio of differentials on the left can be interpreted as the slope of the contour at a point, since they meet the restriction that their values are such as to leave the value of the function unchanged. Thus [B.9] shows that we can evaluate the slope of the contour at a point such as x' in Fig. 2.4(a) directly from the values of the partial derivatives of the function at that point. Equation [B.9] also allows us to determine the direction of slope of the contour from the signs of the derivatives of the function. If the derivatives have the same sign, the slope of the contour must be negative as in (a) of Fig. 2.4; while if they have opposite signs, the contour must have a positive slope, as in (b) of the figure.

In optimization theory the continuity of its contours is an important general property which any given function may or may not possess. A second very important property is that of the *concavity* of contours. To examine this, we first restrict ourselves to functions whose derivatives f_1 and f_2 are positive; it follows from this that given two vectors x' and x , $x > x' \Rightarrow f(x) > f(x')$. Figure 2.5 illustrates concavity of contours for such a function. The property can be described as follows: choose two points, such as \bar{x} and x'' in the figure, which lie on the same contour. That is, $f(x') = f(x'') = c$. Choose any point

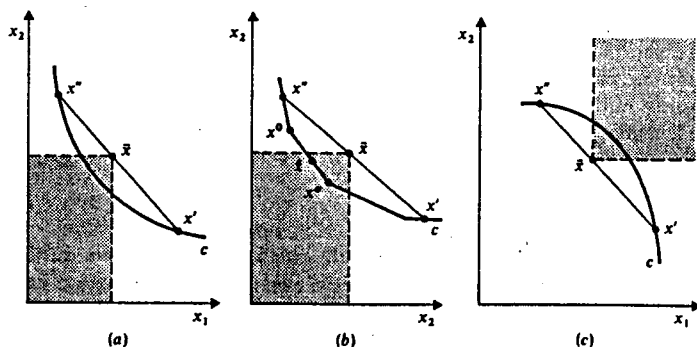


Fig. 2.5

on the straight line joining x' and x'' , such as \bar{x} in the figure. Then, the contour is said to be concave if:

$$f(\bar{x}) \geq f(x') = f(x'') = c \quad [\text{B.10}]$$

In words, a convex combination of any two points on a contour yields at least as high a value of the function and so lies on the same or a higher contour. A function $f(x)$ is *quasi-concave* if

$$f(x') \geq f(x'') \text{ implies } f(kx' + (1-k)x'') \geq f(x'') \quad 0 \leq k \leq 1 \quad [\text{B.11}]$$

when $f(x') = f(x'')$ [B.11] yields [B.10] so that quasi-concave functions have concave contours. The functions whose contours are shown in (a) and (b) of Fig. 2.5 are quasi-concave,⁴ while that in (c) is not. To see this, note that in (a) and (b), part of each contour passes through the shaded area southwest of \bar{x} , implying that some points on the contour have smaller values of both x_1 and x_2 than at \bar{x} . Since decreasing x_1 and x_2 must reduce the value of the function ($f_1, f_2 > 0$), it follows that the value of the function must be smaller on the contour than at \bar{x} , and so [B.10] is satisfied. In (c) on the other hand, this is not the case; part of the contour lies in the area northeast of \bar{x} , and so contains points at which both x_1 and x_2 are greater than at \bar{x} . Hence, the value of the function is greater along the contour.

A further distinction can be drawn by considering (a) and (b) of the figure. In (a), it is clear that for any two points on the contour, the line joining them will always lie wholly above the contour, implying that the value of the function at a point such as \bar{x} will always be strictly greater than that on the contour. Such a function is called *strictly quasi-concave*. This is not, however, true for the contour in (b). For example, if we take points x_0 and x^* on the contour, then a point such as \bar{x} on the line joining them also lies on the contour, and so:

$$f(\bar{x}) = f(x^*) = f(x_0) = c \quad [\text{B.12}]$$

Hence, a function possessing such contours, though quasi-concave, is not *strictly* quasi-concave.⁵

Note, finally, that changing the value of the constant c will change the contour of the function. Given $f_1, f_2 > 0$, increasing c will shift the contours in Fig. 2.5 rightward, since for any given value of x_2 , a greater value of x_1 will be required to satisfy the equation defining the contour (and conversely). It follows that in an optimization problem the higher the contour attained, the greater the value of the objective function, so that we can regard the aim of maximizing the objective function as equivalent to getting onto the highest possible contour.

Properties of the feasible set

There are four important properties of point sets which are of interest in optimization theory.

1. Non-emptiness. A set is non-empty if it contains at least one element, the empty set being the set with no elements. Recall that the feasible set in a problem is the set of points or vectors x which satisfy the constraints. An empty feasible set implies that no such points exist: the constraints are such as to rule out all possible solutions. If the constraints can be satisfied by at least one point, the feasible set is non-empty.

2. Closedness. A set is closed if *all* the points on its boundaries are elements of the set. Thus the set of numbers x on the interval $0 \leq x \leq 1$ is closed, while those sets defined on the intervals $0 < x \leq 1$ and $0 \leq x < 1$, are not. As a later exercise shows, there is a close relationship between the existence of *weak* inequalities in the constraints of a problem, and the closedness of the feasible set.

3. Boundedness. A set is bounded when it is not possible to go off to infinity in any direction while remaining within the set. In other words, it will always be possible to enclose a bounded set within a sphere of sufficiently large size. Thus the set of numbers x on the interval $0 < x < 1$ is bounded, while the set $x \geq 0$ is unbounded. Note that boundedness and closedness are quite distinct: the set defined by $0 < x < 1$ is bounded but not closed; the set of values $x \geq 0$ is unbounded and closed.

4. Convexity. A set is convex if *every* pair of points in it can be joined by a straight line which lies entirely within the set. If two points in the set can be found such that the line joining them lies at least in part outside the set, then the set is non-convex. More formally, a set X is convex if x, x' are any two points in X and $\bar{x} = kx + (1 - k)x' \in X, 0 \leq k \leq 1$. In Fig. 2.6, (a) shows a number of convex sets, and (b) a number of non-convex sets. If, when *any* two boundary points of a convex set are joined with a line, the whole of the line except its end points is in the interior of the set, then the set is *strictly convex*. If two points can be found such that the line joining them coincides at least in part with the boundary, then the set is not strictly convex. In Fig. 2.6(a), the sets B and C are strictly convex, while A is not.

Recall that the *intersection* of two sets A and B , written $A \cap B$, is the set of points that are in *both* A and B . It is an important fact that the intersection of convex sets is itself a convex set. To prove this, let x, x' be in both A and B , which are convex sets. Then the point $\bar{x} = kx + (1 - k)x', 0 \leq k \leq 1$, is in A and B , because both sets are convex, and so is in $A \cap B$. But since x and x' are any two points in $A \cap B$ this gives the result. Note that this extends easily to any number of convex sets.

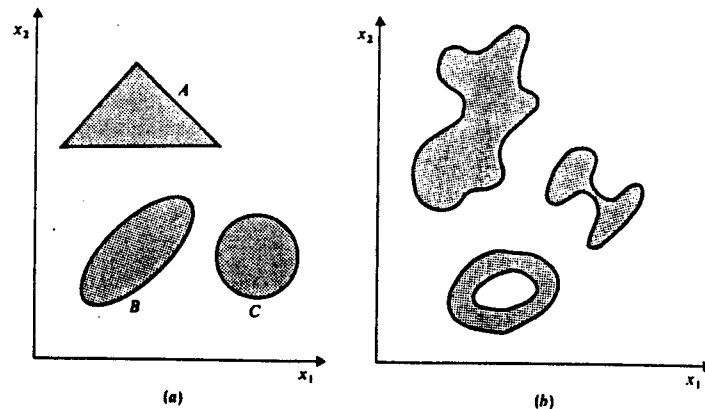


Fig. 2.6

Exercise 2B

- Prove that any global maximum must also be a local maximum.
 - Prove that if for a given problem more than one global maximum exists, the value of the objective function must be the same at each.
 - Show that a linear function is both convex and concave, but neither strictly convex nor strictly concave.
- For each of the cases:
 - $f_1, f_2 < 0$
 - $f_1 > 0, f_2 < 0$
 - $f_1 < 0, f_1 > 0$
 where f_1 and f_2 are the partial derivatives of the function $f(x_1, x_2)$, define and draw contours of quasi-concave, strictly quasi-concave and non-quasi-concave functions. Indicate in each case the direction in which higher contours are attained.
- Given some point x' on a contour of the function $f(x)$, define the 'better set' B' as that set of points which have the property: $f(x) \geq f(x')$. Can you frame definitions of quasi-concavity and of strict quasi-concavity of the function in terms of a property of this set B' ?
- Using the definitions in equations [B.4] and [B.10] or [B.11], show that a concave function is always quasi-concave. To prove that the converse need not hold, sketch or describe in three dimensions a quasi-concave non-concave function.
- Take each of the feasible sets of Question 2, Exercise 2A, and state whether it is non-empty, closed, bounded and convex.
- Explain why a set may be unbounded but closed, or bounded but not closed.
- Recall from [B.9] the expression for the slope of a contour. The curvature of the contour is determined by the second derivative d^2x_1/dx_1^2 . The function is strictly quasi-concave if $d^2x_1/dx_1^2 > 0$. Show by differentiating through [B.9] that this requires

$$\frac{1}{f_1^2} \{ f_1^2 f_{22} - 2f_1 f_2 f_{12} + f_2^2 f_{11} \} < 0$$

C. Existence of solutions

Armed with these intuitive ideas about objective functions and the feasible set, we can now try to answer some of the questions posed at the beginning of Section B. The most fundamental question which can be asked of an optimization problem is whether a solution exists. We can specify conditions on the properties of the feasible set and the objective function in an optimization problem which ensure that there is a solution to that problem. These conditions are embodied in *Weierstrass' Theorem*:

An optimization problem always has a solution if:

- the objective function is continuous; and the feasible set is:
- non-empty

- (c) closed, and
(d) bounded

This theorem is based on the fact that the set of values of the function $f(x)$, which results when we plug into it the x -values in the feasible set, is itself non-empty, closed and bounded, given that the conditions of the theorem are satisfied. That is, a continuous function maps or transforms a closed and bounded set of vectors onto a closed and bounded set of real numbers. Any such set contains its lowest upper bound, which is the maximum of the function over the set, and its greatest lower bound, which is therefore the minimum.

A very simple illustration of the roles played by the conditions in the theorem is given in Fig. 2.7. We take a one-variable problem, where the objective function is given by $f(x)$, x a scalar, and the feasible set by the set of values on the interval $0 \leq x \leq x'$. This feasible set is non-empty, closed, and bounded. In (a) of the figure the function $f(x)$ is continuous and a solution to the problem of maximizing f over the feasible set is found at x' , the upper boundary of the feasible set. In (b) on the other hand, the function is discontinuous at x^0 . In that case there is no solution to the maximization problem; by letting $x \rightarrow x^0$, we can go on increasing the value of the function, since $\lim_{x \rightarrow x^0} f = \infty$. The condition of continuity rules out such cases as (b).

To see the importance of closedness, suppose that the feasible set is defined by the interval $0 \leq x < x'$, so that the upper boundary x' is not in the set. Then, in (a), if we let $x \rightarrow x'$, we can go on increasing the value $f(x)$ without end, since we can let x get closer and closer to x' without ever attaining it. In other words y_{\max} is not an element of the set of values of $f(x)$. Thus the maximization problem has no solution.

Boundedness is important because in its absence we again have the possibility that the value of the objective function can be made to increase without limit. Thus suppose that the feasible set in the problem is given simply by the constraint $x \geq 0$, and also that the objective function is monotonically increasing for $x > x'$ in Fig. 2.7(a). Then clearly there will be no maximum. Boundedness of the feasible set rules out this kind of case.

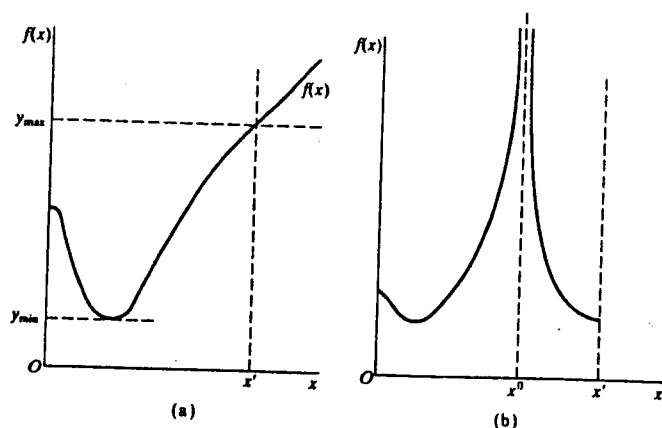


Fig. 2.7

It should be noted that the conditions of continuity of the objective function and closedness and boundedness of the feasible set are *sufficient but not necessary* conditions for existence of a solution. In other words, solutions *may* exist if they are not satisfied, but solutions may also not exist. Satisfaction of the conditions, however, rules out all possible cases of non-existence. Note, finally, that the condition of non-emptiness of the feasible set is a necessary condition for existence of a solution; any problem in which no point is feasible cannot have a solution.

Exercise 2C

1. Draw variants of Fig. 2.7 which show that solutions may exist when the objective function is discontinuous, and the feasible set open and unbounded. From this, explain what is meant by the statement that the conditions of the existence theorem are sufficient but not necessary.
2. Explain why the only condition of the existence theorem which is necessary is that the feasible set be non-empty.

D. Local and global optima

Suppose that the conditions which guarantee existence of a solution are satisfied; we have a continuous objective function, and a non-empty, closed and bounded feasible set. We consider a two-variable problem, in which we wish to maximize the function $f(x_1, x_2)$, with $f_1, f_2 > 0$. We are now interested in the question: given that we can find a local maximum of this function, under what conditions can we be sure that it is also a global maximum?

Two possibilities, from which we can abstract the conditions we seek, are set out in Fig. 2.8, where the shaded areas are the feasible sets. Recall that maximization of a function over a given feasible set is equivalent to finding a point within that set which is on the

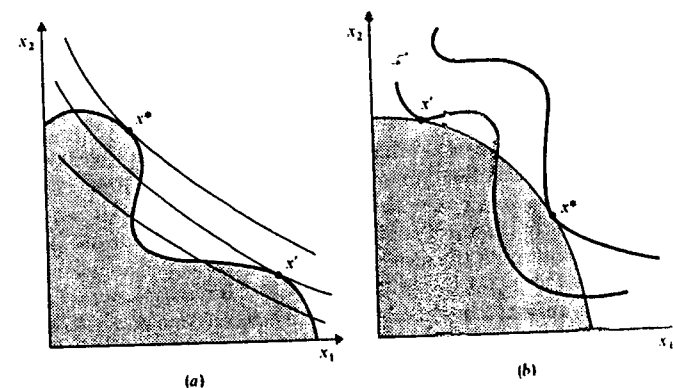


Fig. 2.8

highest possible contour. Given the assumption $f_1, f_2 > 0$, contours increase in value as we move northeastwards in the figure. Then both (a) and (b) show examples of cases in which two local optima exist, only one of which is a global optimum. In (a) the objective function is strictly quasi-concave and the feasible set is non-convex. There is a local maximum at x^* and also at x' , since relative to a small neighbourhood of points within the feasible set around them, they are on the highest possible contours. By inspection, we can see that x^* only is a global maximum. In (b), the feasible set is convex, and the objective function is not quasi-concave: there are local optima at x^* and x' , but only x^* is a global maximum.

Consider now the problem of framing sufficient conditions for any local optimum also to be global. Figure 2.8 helps us to see intuitively that they must depend on the shapes of the feasible set and of the contours of the function. As (a) shows, it is not sufficient that the function be quasi-concave; and (b) shows that it is not sufficient that the feasible set be convex. However, taking both together, we can state the theorem:

A local maximum is always a global maximum if:

- (a) *the objective function is quasi-concave, and*
- (b) *the feasible set is convex.*

Proof: The proof is by contradiction. Let x^* be a local maximum, i.e. $f(x^*) \geq f(x)$ for all x in a neighbourhood of x^* , and suppose there exists $x' \in S$, the feasible set, such that $f(x') > f(x^*)$. Then since S is a convex set we have

$$\bar{x} = kx^* + (1 - k)x' \in S \quad \text{for } 0 \leq k \leq 1$$

In addition, since f is quasi-concave and $f(x') > f(x^*)$ we have

$$f(\bar{x}) \geq kf(x^*) + (1 - k)f(x') > f(x^*)$$

But then, by choosing k arbitrarily close to 1, we can ensure that \bar{x} is in a neighbourhood of x^* , contradicting the fact that x^* is a local maximum. Thus if x^* is a local maximum there cannot exist $x' \in S$ such that $f(x') > f(x^*)$ and so x^* is also a global maximum.

To get some intuitive understanding of this theorem consider Fig. 2.9. The feasible set S is convex, and the objective function, with contour c , is quasi-concave. This latter implies that the set B , consisting of those points along and above the contour c , is also convex. The solutions to the problem consist of the points on the segment ab , since these are the points in S on the highest possible contour. Since these points lie on the same contour they yield the same value of the objective function; each is a global as well as a local maximum. Now consider the line T , part of which is coincident with the segment ab . Because S is a convex set, and the segment ab lies along its upper boundary, the set S must lie on or below T – no point in S can lie above T . Likewise, because B is a convex set (as a result of the quasi-concavity of the objective function) and ab lies along its lower boundary, the entire set B must lie on or above T – no point in B can lie below T . But B is the set of points which yield at least as high a value of the objective function as the points along ab . Thus we have the conclusion that no points in B can possibly also be in

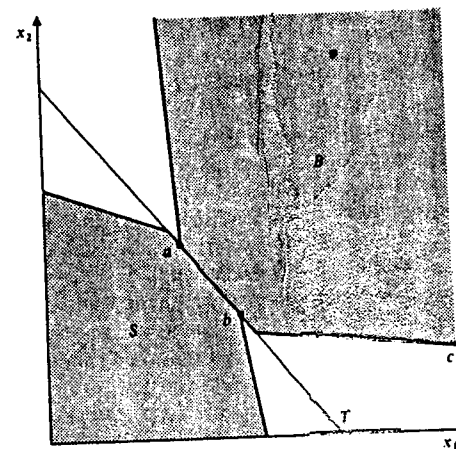


Fig. 2.9

S except that along the segment ab . But this in turn implies that the points along ab must be global as well as local optima.

The difference between the cases in Figs. 2.8 and 2.9 is that in the former the absence of convexity or quasi-concavity implies that the set B may intersect with the set S at points other than a given local optimum, and so that optimum may not be global. Convexity of the sets B and S rules this out. In this case the line T is said to separate the sets S and B , and the importance of the convexity of B and S lies in the fact that such a separating line can always be found. Note that, again, the conditions of the theorem are sufficient but not necessary: even if they do not hold, the configuration of contours and feasible set may be such that local optima are also global.

Finally, since a concave function is always quasi-concave (see Question 4, Exercise 2B), we can conclude that the theorem also holds for concave functions.

Exercise 2D

1. (a) Draw examples of cases in which the conditions of the theorem are not met, but local optima are also global optima.
(b) Explain why the set B in Fig. 2.9 consists of points which yield at least as high values of the objective function as those along the segment ab .
2. Suggest what happens to the solution illustrated in Fig. 2.9 if the feasible set S is not closed.
- 3.* Consider the adaptation of the theorem of this section to the case in which it is desired to minimize the function $f(x_1, x_2)$ (Hint: review the definition of quasi-convexity in note 5).
4. Adapt the discussion of Fig. 2.9 to the case in which the set B is strictly convex.

E. Uniqueness of solutions

From a normative or prescriptive point of view the question of the uniqueness of solutions is not very important: by definition one global optimum is as good as another. However, if we are using the optimization problems for positive or predictive purposes, the question of whether the decision-maker has a unique best decision or a number of equally good decisions is of more relevance. We are interested in the way in which decisions change in response to changes in the constraints defining the feasible set. Where the optimal solution is unique for each given feasible set we can specify functions which relate the optimal values of the choice variables to the parameters in the constraints. For example, this is how we derive demand, cost and supply functions in economics. If, on the other hand, the solution is not unique, then we have a more general relationship between optimal *sets* of values of the choice variables and the constraint parameters, known as a *correspondence*. Though this presents no insuperable obstacles to analysis, it does require us to change our procedures and approach. Since economics (at the level covered by this book) usually deals with functions rather than with correspondences, it is worthwhile to be aware of the precise circumstances in which it is valid to do this. These circumstances exist when the solutions to the relevant optimization problems are unique.

In Fig. 2.9 of the previous section we saw a case in which there were multiple global optima; there was an infinite number of optimal points on the line segment ab . In that case the feasible set was convex, but not strictly convex; and the objective function was quasi-concave, but not strictly quasi-concave. Consider now Fig. 2.10, where we show three unique global optima, respectively x^* , x' and x'' . In the first the objective function is strictly quasi-concave but the feasible set not strictly convex; in the second we have the reverse; and in the third the objective function is strictly quasi-concave and the feasible set strictly convex. These figures illustrate the *Uniqueness Theorem*:

Given an optimization problem in which the feasible set is convex and the objective function is non-constant and quasi-concave, a solution is unique if:

- (a) the feasible set is strictly convex, or
- (b) the objective function is strictly quasi-concave, or
- (c) both

Proof: The proof is by contradiction. Let x^* be a solution and suppose there exists $x' \in S$, the feasible set, such that $x' \neq x^*$ and $f(x') = f(x^*)$. Since S is convex, it contains $\bar{x} = kx^* + (1-k)x'$, $0 \leq k \leq 1$. Now if f is strictly quasi-concave we have immediately from the definition

$$f(\bar{x}) > kf(x^*) + (1-k)f(x') = f(x^*)$$

which contradicts the optimality of x^* . Therefore in this case the optimum must be unique. Now suppose the feasible set is strictly convex. Then \bar{x} lies in the interior of S and, since $f(x)$ is non-constant in x , it is always possible to find another point in S which yields a higher value of the function. Again this contradicts the optimality of x^* .

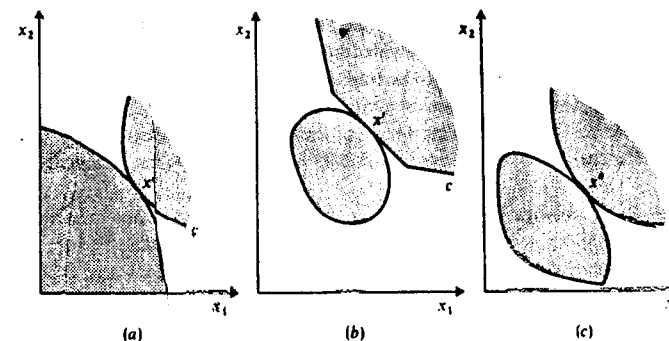


Fig. 2.10

As usual, the theorem gives only sufficient conditions: it is possible that a unique solution will exist even when the conditions are not satisfied, but we cannot be sure.

Exercise 2E

1. Draw examples in which the optimum is unique even though the conditions of the theorem are not met.
2. Apply the theorem of this section to the case of minimization problems (refer back to Question 3, Exercise 2D).
- 3.* When $f_1, f_2 > 0$, show that it is sufficient for uniqueness that the feasible set be upper convex, i.e. a line joining any two points on its upper boundary lies in the interior of the set. (Hint: consider Fig. 2.10(a).)

F. Interior and boundary optima

The boundaries of a feasible set are always defined by the constraints which are part of the specification of the optimization problem. Given that the feasible set is closed, we can partition its points into two mutually exclusive subsets, interior points and boundary points. Loosely, the former points lie inside the boundaries of the set and the latter lie on them. More rigorously, the defining characteristic of an interior point of a point set is that we can find a (possibly very small) neighbourhood around it which contains *only* points in the set. A boundary point, on the other hand, has the property that *all* neighbourhoods around it, however small, contain points which are and points which are not in the set (draw diagrams to show that these definitions are consistent with your intuition).

Recall why we are interested in the distinction between interior and boundary points. In general, a solution to an optimization problem which is at an interior point of a feasible set is unaffected by small shifts in the boundaries of the set, while a solution at a boundary point will be sensitive to changes in at least one constraint. Since much of microeconomics is concerned with predicting the changes in solutions to optimization

problems resulting from shifts in constraints, the question of whether such solutions are at interior or boundary points is fundamental.

In parts (a) and (b) of Fig. 2.11 the feasible sets are initially the areas Oab . In (a) we have an interior optimum at x^* , and in (b) and (c) we have boundary optima also denoted x^* . The solution in (a) is unaffected by a small shift in the constraint, e.g. to $a'b'$; that in (b) is affected; that in (c) is changed by a shift in constraint cd but not by that in ab , as illustrated.

The absence of response of the solution in (a) is due to the assumed existence of a bliss point at x^* (the 'peak' of the 'hill' whose contours are drawn in the figure), i.e. a point at which the objective function takes on a maximum. The occurrence of a bliss point in the interior of the feasible set is clearly necessary for there to be an interior maximum and so we can characterize cases of boundary maxima as ones in which no bliss points exist (but see Question 1, Exercise 2F). One such class of cases is that in which the objective function is monotonically increasing, i.e. every $f_i > 0$, where f_i is the i th partial derivative of the function. In these cases we can go further and say that the solution must be on the upper boundary of the feasible set. In terms of the contours of the function, this would imply that higher contours are reached as we move rightwards in the diagram. Clearly, however, it is not necessary, if we want to rule out bliss points, that the partial derivatives are all positive or even that the objective function is differentiable. It is simply necessary to assume that at any point in the feasible set it is always possible to find a small change in the value of at least one variable which will increase the value of the objective function. This is the property of *local non-satiation*.

Parts (b) and (c) of the figure show two kinds of boundary optima. In (b) there is only one upper boundary and, given the assumption that $c_2 > c_1$, the boundary shift changes the optimum. In (c), the initial feasible set is taken to be the area $Oceh$ defined by two weak linear inequalities and non-negativity constraints on x_1 and x_2 . The initial optimum is on the boundary at x^* . At such a point the constraint defined by the line ab is satisfied as a strict inequality. This constraint is inoperative at the solution and so is non-binding. Once we know where the solution lies, a non-binding constraint can be dropped from any analysis concerned with movements around a small neighbourhood of the optimum, and this can often greatly simplify such analysis. Before solving the problem, however, it is impossible to know which constraints will turn out to be non-binding (given that we have

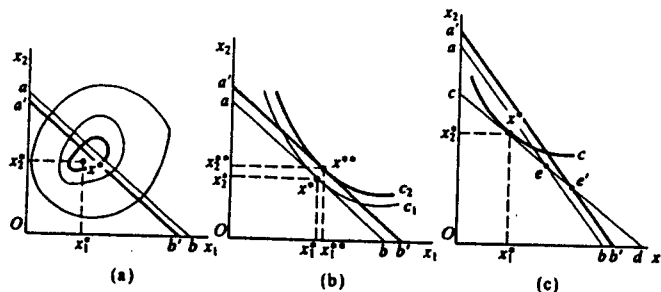


Fig. 2.11

eliminated constraints which lie wholly outside other constraints and so could not possibly be binding) and so all must be retained. Moreover, in a general theoretical analysis we do not have enough information to conclude that some constraint will turn out to be non-binding and so all solution possibilities have generally to be considered. In Fig. 2.11 (a) there are three such solution possibilities – point out the two not shown. What can be said in the latter two cases about the responsiveness of the solution to shifts in the constraints?

Exercise 2F

1. Draw an example of a case in which a bliss point exists, but the solution is affected by some kinds of constraint shifts.
- 2.* The theory of 'satisficing' says that given a feasible set, an individual chooses not the best alternative but a 'satisfactory' one. Explain why, in terms of the discussion of this section, this theory may fail to yield predictions about economic behaviour.
3. Given a feasible set such as that in Fig. 2.11 (a), where would you expect the solution to be in the cases:
 - (a) $f_1 > 0, f_2 < 0$
 - (b) $f_1 < 0, f_2 > 0$
 - (c) $f_1, f_2 < 0$
 - (d) $f_1 = 0, f_2 > 0$
 where $f(x_1, x_2)$ is the objective function and $f_i, i = 1, 2$ its partial derivatives.
- 4.* If all consumers in an economy possessed bliss points, what would be the relevance of microeconomics?

G. Location of the optimum: the method of Lagrange

As suggested earlier, in general theoretical models we do not have numerical information with which to find solutions to optimization problems. Instead, we seek to describe the characteristics or properties the solution possesses in terms of general conditions which it satisfies.

It is assumed that the reader is familiar with the necessary condition for the point $x^* = (x_1^*, x_2^*, \dots, x_n^*)$ to yield a maximum of the function $f(x)$ when n constraints are present, namely

$$f_i(x^*) = 0 \quad i = 1, 2, \dots, n \quad [G.1]$$

that is, that each partial derivative of the function, evaluated at x^* , must be zero.

Suppose now, however, we also have a constraint on the problem: we seek the vector giving the largest possible value of $f(x)$ from only those vectors which satisfy the constraint $g(x) = b$, where g is a differentiable function and b is a real number, the *constraint constant*.

We then write the problem as

$$\max_x f(x) \text{ s.t. } g(x) = b \quad [\text{G.2}]$$

Referring back to parts (a) and (b) of Fig. 2.11 allows us to see immediately that the conditions [G.1] are no longer necessary for a solution to [G.2]. In Fig. 2.11(a), conditions [G.1] are satisfied at the bliss point x^* , the 'peak of the hill', but if we are constrained to choose a solution *upon* the line ab in the figure, in which case the constraint in [G.2] takes the linear form

$$a_1 x_1 + a_2 x_2 = b \quad [\text{G.3}]$$

then this solution will clearly lie at a point on the 'side' of the hill where *some* $f_i \neq 0$. Similarly, at point x^* in Fig. 2.11(b), which is a solution to the problem again with a linear constraint ab , the partial derivatives f_i cannot possibly be zero, because we have (recalling [B.8] and [B.9])

$$\frac{dx_2}{dx_1} = -\frac{f_1}{f_2} = -\frac{a_1}{a_2} \quad [\text{G.4}]$$

where $-a_1/a_2$ is the slope of the linear constraint [G.3].

So, a point which solves the problem [G.2] need not (and generally will not) satisfy conditions [G.1], which are then no longer necessary conditions for a solution. We have to develop a new set of conditions. We do so by extending the idea just introduced in [G.4]. To do this, suppose that the problem in [G.2] can be represented as in Fig. 2.12. We now assume $g(x_1, x_2)$ is non-linear, and $g(x_1, x_2) = b$ defines a contour of this function. Similarly, $f(x_1, x_2) = c$ defines a contour of the objective function, and the optimum is at x^* . Note that we have built into the diagram the assumptions which ensure that x^* exists, and is a unique global optimum.

The essential fact about x^* in the figure is that it is a point of tangency. That is, at x^* , the contour of the objective function f and the contour of the constraint function g have a slope equal to that of the common tangent T . We have already shown (see equation [B.9]) that the slope of any contour of f is given by:

$$\frac{dx_2}{dx_1} = -\frac{f_1}{f_2} \quad [\text{G.5}]$$

By a similar argument (supply the details) we can show that the slope of the constraint contour is:

$$\frac{dx_2}{dx_1} = -\frac{g_1}{g_2} \quad [\text{G.6}]$$

It follows that the optimal solution x^* satisfies the conditions:

$$(i) \quad \frac{f_1}{f_2} = \frac{g_1}{g_2} \quad [\text{G.7}]$$

$$(ii) \quad g(x_1^*, x_2^*) = b$$

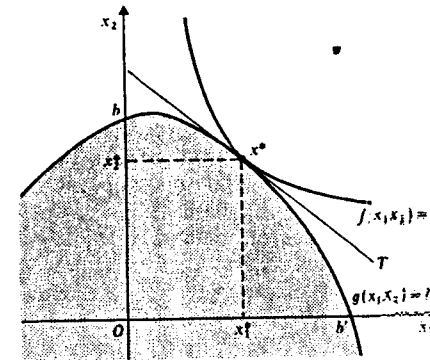


Fig. 2.12

Note that (ii) is an important part of these conditions. (i) simply states that we must be at *some* point of tangency, but does not precisely locate the optimum point – the single equation cannot determine values of both x_1 and x_2 . The addition of (ii) closes the system, and ensures that what we have is actually the point of tangency on the constraint contour. (Note that the condition $f(x_1^*, x_2^*) = c$ would do just as well provided we know the value of c at the optimum.)

(i) can be expressed as:

$$\frac{f_1}{g_1} = \frac{f_2}{g_2} = \lambda^* > 0 \quad [\text{G.8}]$$

where λ^* is simply a number representing the common value of the ratios f_i/g_i , $i = 1, 2$, at the optimum. But [G.8] then implies the two equations:

$$f_1 = \lambda^* g_1 \quad f_2 = \lambda^* g_2 \quad [\text{G.9}]$$

which are then logically equivalent to (i) of [G.7]. Given that $\lambda^* \neq 0$ and $g_i(x_1^*, x_2^*) \neq 0$, $i = 1, 2$, [G.9] implies that at the optimum $f_i \neq 0$, $i = 1, 2$. Thus, as we conjectured, the conditions in [G.1] are *not* necessary for a *constrained* maximum. The optimum x^* in Fig. 12 has been shown, in [G.9] and [G.7], to satisfy conditions which can be written as:

$$\begin{aligned} f_1 - \lambda^* g_1 &= 0 \\ f_2 - \lambda^* g_2 &= 0 \\ g(x_1^*, x_2^*) - b &= 0 \end{aligned} \quad [\text{G.10}]$$

As is always the case with geometrical reasoning, we have built into the analysis a number of very restrictive assumptions simply by drawing a particular picture. In particular we have assumed only two-choice variables, no non-negativity conditions, and only one functional constraint. However, the way of writing the necessary conditions in [G.10] suggests a major step in generalizing the results, by means of a procedure first formulated by the mathematician J. L. Lagrange. Given the problem illustrated in Fig. 2.12, we

formulate the Lagrange function:

$$L(x_1, x_2, \lambda) = f(x_1, x_2) - \lambda[g(x_1, x_2) - b] \quad [\text{G.11}]$$

If we now carry out the unconstrained maximization of the Lagrange function L with respect to its three variables x_1 , x_2 and λ , we obtain precisely the necessary conditions for an optimum shown in [G.10]. (Note that for this we have to multiply the third partial derivative by -1 . To avoid this we could have formed the Lagrange function by taking $g(x_1, x_2)$ over to the righthand side of the constraint, multiplying by λ , and adding to f . Confirm that nothing significant is affected thereby. You will encounter both methods.)

One way of regarding Lagrange's procedure is as follows. We are initially faced with a problem we do not know how to solve, that of maximizing f subject to a constraint. However, we *do* know how to solve unconstrained problems, and so the trick is to turn the constrained problem into an unconstrained one. This is precisely what Lagrange's procedure achieves.

Formal analysis of Lagrange's method*

We need to consider Lagrange's method a little more rigorously. Note first that we could have derived conditions [G.7] by algebraic rather than geometrical reasoning. Assume it is possible to solve the constraint $g(x_1, x_2) - b = 0$ for x_1 as a function of x_2 , say $x_1 = h_1(x_2)$. Then substituting for x_1 in the objective function gives $f(h_1(x_2), x_2) \equiv \Phi(x_2)$. In effect, Φ gives the values of f along the contour in (x_1, x_2) - space defined by $g(x_1, x_2) = b$. Then for a maximum of Φ we must have

$$\Phi'(x_2^*) = f_1 h_1'(x_2^*) + f_2 = -f_1 \frac{g_2}{g_1} + f_2 = 0 \quad [\text{G.12}]$$

where we have used [G.6] to obtain $h_1' = dx_1/dx_2 = -g_2/g_1$. Then [G.12] gives us the 'tangency condition' (i) of [G.7], and the rest of the derivation of the Lagrange procedure follows.

The key assumption here was that we could solve the constraint for one variable as a function of the other, i.e. the implicit function $g(x_1, x_2) - b = 0$ yields $h_1(x_2)$. This suggests that the key theorem underlying Lagrange's method is the *Implicit Function Theorem*. To state this in its general form, suppose we have a system of m continuously differentiable implicit functions in $n > m$ variables, of the form

$$g^1(x_1, \dots, x_m; x_{m+1}, \dots, x_n) = 0$$

$$g^2(x_1, \dots, x_m; x_{m+1}, \dots, x_n) = 0$$

$$\dots$$

$$g^m(x_1, \dots, x_m; x_{m+1}, \dots, x_n) = 0$$

Define D as the $m \times m$ determinant formed by taking the first m partial derivatives of these m functions, i.e.

$$\begin{vmatrix} g_1^1 & g_2^1 & \dots & g_m^1 \\ g_1^2 & g_2^2 & \dots & g_m^2 \\ \dots & \dots & \dots & \dots \\ g_1^m & g_2^m & \dots & g_m^m \end{vmatrix}$$

Then, the Implicit Function Theorem states:

If $D \neq 0$, then we can find m continuous functions $h_j(x_{m+1}, \dots, x_n)$ such that $x_j = h_j(x_{m+1}, \dots, x_n)$ and $g^j(h_1, \dots, h_m; x_{m+1}, \dots, x_n) = 0$, $j = 1, \dots, m$.

Thus the theorem says that a sufficient condition to be able to 'solve' the m implicit functions for m of the variables as functions of the remaining $n-m$ variables is that the determinant D be non-zero.

To obtain some insight into how this theorem relates to Lagrange's method, we apply it first to the simple case of 1 constraint and 2 variables. For the procedure resulting in [G.12] to work, we require $g_1 \neq 0$, since otherwise g_2/g_1 is undefined and [G.12] could not hold. In this simple case g_1 is in fact the determinant D . Suppose, contrary to the condition of the theorem, that $g_1 = 0$. Geometrically, this implies that the constraint contour $g(x_1, x_2) = b$ is a horizontal line in the $x_1 x_2$ -plane, since variations in x_1 leave g unchanged. This gives some intuition as to why we cannot find a function $x_1 = h_1(x_2)$ in this case. Associated with the value of x_2 at the intercept of the horizontal line is an infinity of x_1 -values, while for a function $h_1(x_2)$ to exist there must be only one x_1 -value.

Furthermore, when $g_1 = 0$ and assuming $g_2 \neq 0$ (otherwise the problem is trivial), the Implicit Function Theorem tells us how to proceed. We can solve $g(x_1, x_2) - b = 0$ for x_2 as a function of x_1 , $x_2 = h_2(x_1)$ with $dx_2/dx_1 = h_2'(x_1) = -g_1/g_2 = 0$. Thus $h_2(x_1)$ is a constant function, since values of x_1 map into the same x_2 point, and the graph of the function is the horizontal constraint line. The counterpart of [G.12] is now

$$\Phi'(x_1^*) = f_1 + f_2 h_2'(x_1^*) = f_1 - f_2 g_1/g_2 = f_1 = 0 \quad [\text{G.13}]$$

since $g_1 = 0$. Thus the solution of the problem, if it exists, is the point at which a contour of the objective function is tangent to the horizontal constraint line, i.e. at which its slope $-f_1/f_2 = 0$. Note that such a point may not exist, if, say, $f_1 > 0$ for all x . This results because we have a violation of one of the conditions of Weierstrass' Theorem: the feasible set - the horizontal constraint line - is unbounded. However, if a solution *does* exist, [G.13] tells us that the Lagrange method will find it. Since $g_1 = 0$, the Lagrange conditions will be

$$f_1 = 0$$

$$f_2 - \lambda^* g_2 = 0$$

$$g(x_1^*, x_2^*) - b = 0$$

and $\lambda^* = f_2/g_2$ is well defined. Thus, provided the Implicit Function Theorem holds when applied to the constraint function, the Lagrange method can be applied.

We can now generalize this. Suppose the problem is

$$\max_x f(x) \text{ s.t. } g^1(x) - b_1 = 0, g^2(x) - b_2 = 0, \dots, g^m(x) - b_m = 0 \quad [\text{G.14}]$$

where $x = (x_1, x_2, \dots, x_n)$ and $m < n$. We now introduce a vector of m Lagrange multipliers, $\lambda = (\lambda_1, \dots, \lambda_m)$, one for each constraint, and form the Lagrange function

$$L(x, \lambda) = f(x) - \sum_{j=1}^m \lambda_j [g^j(x) - b_j] \quad [\text{G.15}]$$

Carrying out the unconstrained maximization of $L(x, \lambda)$ yields the conditions

$$\frac{\partial L}{\partial x_i} = f_i - \sum_{j=1}^m \lambda_j^* g_j^i = 0 \quad i = 1, \dots, n \quad [G.16]$$

$$-\frac{\partial L}{\partial \lambda_j} = g^j(x^*) - b_j = 0 \quad j = 1, \dots, m \quad [G.17]$$

Thus we have the generalization of the earlier procedure to n variables and m constraints (note that the restriction $n > m$ is to avoid cases in which the feasible set is either empty or contains only one point). We then claim that the x^* which (along with the λ^*) satisfies these conditions is in fact an x -vector which solves the original constrained problem in [G.14]. It is certainly feasible, since it satisfies [G.17]. The condition under which the claim is true is given by the following

Theorem of Lagrange Multipliers: If the constraints in the problem [G.14] are written in the form

$$g^j(x_1, x_2, \dots, x_m; x_{m+1}, \dots, x_n) - b_j = 0 \quad j = 1, \dots, m$$

and the determinant

$$D = \begin{vmatrix} g_1^1 & g_1^2 & \dots & g_1^m \\ g_2^1 & g_2^2 & \dots & g_2^m \\ \dots & \dots & \dots & \dots \\ g_m^1 & g_m^2 & \dots & g_m^m \end{vmatrix}$$

is non-zero, then there exist m real numbers $\lambda_1, \dots, \lambda_m$ such that the x^* which satisfies [G.16] and [G.17] is the solution to problem [G.14].

A general proof of this theorem is notationally quite complex, but its form can be brought out in a simple example. It uses the same basic ideas we have encountered in the two-variable one-constraint case. Assume now we have three variables and two constraints, $g^j(x_1, x_2, x_3)$, $j = 1, 2$. The Lagrange conditions take the form

$$\begin{aligned} f_1 - \lambda_1 g_1^1 - \lambda_2 g_1^2 &= 0 \\ f_2 - \lambda_1 g_2^1 - \lambda_2 g_2^2 &= 0 \\ f_3 - \lambda_1 g_3^1 - \lambda_2 g_3^2 &= 0 \end{aligned} \quad [G.18]$$

All derivatives are evaluated at the point x^* , the optimal solution to the constrained problem, and so are simply given numbers. Assume the determinant

$$D = \begin{vmatrix} g_1^1 & g_1^2 \\ g_2^1 & g_2^2 \end{vmatrix} = \begin{vmatrix} g_1^1 & g_2^1 \\ g_1^2 & g_2^2 \end{vmatrix} \quad [G.19]$$

is non-zero, so the condition of the theorem (and of the Implicit Function Theorem) is satisfied. Then, from the first two conditions of [G.18] we can solve for λ_1 and λ_2 , using Cramer's rule, to obtain

$$\lambda_1 = \frac{f_1 g_2^2 - f_2 g_1^2}{D}, \quad \lambda_2 = \frac{f_2 g_1^1 - f_1 g_2^1}{D} \quad [G.20]$$

If we can show that these values of λ_1 and λ_2 also satisfy the third condition of [G.18], then we have proved the theorem. From the Implicit Function Theorem, we know that we can use the two constraints to solve for x_1 and x_2 as functions of x_3 , which we can denote $h_1(x_3)$ and $h_2(x_3)$ respectively. Furthermore, from the system

$$\begin{aligned} g_1^1 dx_1 + g_2^1 dx_2 + g_3^1 dx_3 &= 0 \\ g_1^2 dx_1 + g_2^2 dx_2 + g_3^2 dx_3 &= 0 \end{aligned} \quad [G.21]$$

obtained by differentiating the constraints (at x^*), we have that

$$\begin{aligned} h_1'(x_3^*) &= \frac{dx_1}{dx_3} = \frac{g_3^2 g_1^1 - g_3^1 g_2^1}{D} \\ h_2'(x_3^*) &= \frac{dx_2}{dx_3} = \frac{g_3^1 g_2^2 - g_3^2 g_1^2}{D} \end{aligned} \quad [G.22]$$

Substituting for x_1 and x_2 into the objective function to obtain $f(h_1, h_2, x_3) \equiv \Phi(x_3)$, we have that at the optimal point

$$\Phi'(x_3^*) = f_1 h_1'(x_3^*) + f_2 h_2'(x_3^*) + f_3 = 0 \quad [G.23]$$

Then substituting into [G.23] from [G.22] and rearranging terms gives

$$f_3 - g_3^1 \frac{(f_1 g_2^2 - f_2 g_1^2)}{D} - g_3^2 \frac{(f_2 g_1^1 - f_1 g_2^1)}{D} = 0 \quad [G.24]$$

which is precisely the third condition in [G.18], given the solutions in [G.20]. Thus we have the result.

Since in this book we repeatedly apply Lagrange's method, we shall always assume, explicitly or implicitly, that the conditions of the above theorem are satisfied.

Interpretation of the Lagrange multipliers

We derived necessary conditions for an optimal vector x^* by introducing the Lagrange multipliers λ_j and forming the Lagrange function. The multipliers are, however, more than an ingenious mathematical device, and turn out to have an interpretation which is of great interest in specific economic contexts. To show this, we revert to the two-variable problem in Fig. 2.12. Given the necessary conditions for an optimal solution in [G.10], we can regard these as three equations in the three 'unknowns' x_1^* , x_2^* and λ^* , with the constraint value b as an exogenous parameter which determines the solution. We can solve for the unknowns as functions of this parameter, i.e. we may write:

$$\begin{aligned} x_1^* &= h_1(b) \\ x_2^* &= h_2(b) \\ \lambda^* &= h_\lambda(b) \end{aligned} \quad [G.25]$$

which express the idea that the solution values depend upon the parameter b . We define the optimized value, v^* , of the objective function as the value it takes on at the optimal

point, i.e.

$$v^* = f(x_1^*, x_2^*)$$

Clearly, therefore, v^* depends on b , and so we can write, using [G.25];

$$v^* = f(h_1(b), h_2(b)) = v^*(b)$$

Consider the derivative dv^*/db . This gives the rate at which changes in the constraint parameter b cause changes in the optimized value of the objective function, via its effect on the solution values x_1^* and x_2^* . The significance of λ^* stems from the fact that:

$$\frac{dv^*}{db} = \lambda^*$$

In other words, λ^* measures the rate at which the optimized value of the objective function varies with changes in the constraint parameter.

To prove this, note first that, from [G.27]:

$$\frac{dv^*}{db} = f_1 \frac{dx_1^*}{db} + f_2 \frac{dx_2^*}{db}$$

Since all derivatives are evaluated at the optimal point, we have from the conditions [G.10] that $f_1 = \lambda^* g_1$ and $f_2 = \lambda^* g_2$, and so:

$$\frac{dv^*}{db} = \lambda^* \left(g_1 \frac{dx_1^*}{db} + g_2 \frac{dx_2^*}{db} \right)$$

Now since the constraint is satisfied at the optimal point, we have:

$$g(x_1^*, x_2^*) = b$$

so:

$$dg = g_1 dx_1^* + g_2 dx_2^* = db$$

and so:

$$\frac{dg}{db} = g_1 \frac{dx_1^*}{db} + g_2 \frac{dx_2^*}{db} = 1$$

Then, substituting from [G.33] into [G.30] gives the equality in [G.28].

We can interpret this result with the help of Fig. 2.13. Initially, the constraint is defined by $g(x_1, x_2) = b$, and the solution is at x^* . Now suppose the constraint becomes $g(x_1, x_2) = b'$, with $b' > b$, and the curve in the figure shifts outward. There will be a new solution at x^{**} , with optimal values x_1^{**}, x_2^{**} . The change in b has caused a change in the optimized value of the objective function, given by:

$$f(x_1^{**}, x_2^{**}) - f(x_1^*, x_2^*) = v^{**} - v^*$$

Thus, we can take the ratio of these changes:

$$\frac{\Delta v^*}{\Delta b} = \frac{v^{**} - v^*}{b' - b}$$

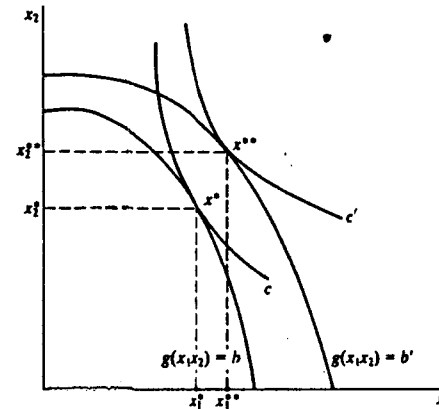


Fig. 2.13

which relates to a finite change in b . Then, in the usual way, we can take:

$$\lim_{\Delta b \rightarrow 0} \frac{\Delta v^*}{\Delta b} = \frac{dv^*}{db}$$

[G.36]

Thus, we can think of the constraint shifting infinitesimally and the derivative dv^*/db then measures the rate at which v^* changes, given that an optimal point is always chosen.

In economics, derivatives are usually designated by the term 'marginal this or that'. The equality in [G.28] implies that λ^* can be thought of as the marginal change in the optimized value of the objective function with respect to changes in the constraint. Then, in specific contexts, this leads to useful interpretations of the Lagrangean multipliers. For example, in the case in which the consumer maximizes utility (the objective function) subject to a budget constraint ($b = \text{income}$), λ^* would measure the *marginal utility of income* at the optimal point. In a problem in which costs were to be minimized subject to a fixed output constraint ($b = \text{output}$), λ^* would measure *marginal cost* at the optimal point. In any problem, there exists an interpretation of the multiplier which is of interest to economists.

A further interpretation of the Lagrange multiplier can be made, as a kind of price. Since its value at the optimum measures the change in value of the objective function caused by a slight shift in the constraint, it can be interpreted as measuring the maximum 'payment' which would be made in exchange for a shift in the constraint. For example, suppose that the problem is to maximize profit by choosing input and output levels, subject to a limitation on the amount of one input available ($b = \text{quantity of input}$). Then λ^* in this problem measures the marginal profitability of the input, or the rate at which maximum profit increases with a small increase in the fixed amount of the input. It follows that λ^* measures the maximum amount the firm would be prepared to pay for the increase in input, since anything less would result in a net increase in profit, and anything more would result in a net decrease. For this reason, λ^* would be called the *shadow price* of the input. The word 'shadow' occurs because it may differ from the input's actual price.

Exercise 2G

1. Sketch in two and three dimensions the case in which a monotonically increasing objective function takes on a maximum over a closed, bounded feasible set. Explain from this why the condition in [G.1] is not necessary for a local maximum in a constrained maximization problem.

- 2.* Explain why the conditions in [G.16] and [G.17] are also necessary for a solution to the problem:

$$\min f(x) \text{ s.t. } g^j(x) - b_j = 0 \quad (j = 1, \dots, m)$$

and why therefore these conditions are necessary but not sufficient for a maximum.

3. Interpret the Lagrange multipliers which would be associated with the constraints in the following optimization problems:

(a) A central planner in a developing country wishes to maximize GNP, subject to the constraints that the balance of payments deficit may not exceed a given figure, and that a fixed amount of skilled labour is available.

(b) A firm wishes to choose a set of investment projects which maximize its profitability, subject to the constraint that the total amount it spends on investment does not exceed a fixed amount of funds available.

4. Consider the problem:

$$\max_{x_1, x_2} f(x_1, x_2) \quad \text{where } f_1, f_2 > 0$$

$$\text{s.t. } a_{11}x_1 + a_{12}x_2 \leq b_1$$

$$a_{21}x_1 + a_{22}x_2 \leq b_2$$

The objective function f is strictly quasi-concave, and the a_{ij} , $i, j = 1, 2$, are all positive.

- (a) Draw the feasible set, assuming

$$\frac{a_{11}}{a_{12}} > \frac{a_{21}}{a_{22}} \quad \text{and} \quad \frac{b_1}{a_{12}} > \frac{b_2}{a_{22}}, \quad \text{and that the lines intersect.}$$

Find the points at which a solution may be found, and suggest their main characteristics, in terms of whether they imply zero or non-zero variable values, and binding or non-binding constraints.

- (b) Recalling the interpretation of Lagrangean multipliers, give an economic interpretation of the case $\lambda_1^* = 0$.

H. Concave programming and the Kuhn-Tucker conditions

In the previous section we considered the problem of maximizing some function subject to functional constraints in strict equality form. Such problems are often referred to as 'classical optimization problems'. Though this type of problem is quite standard in

economics, it involves two important assumptions which are not in general satisfied in many economic problems. First, it assumes that there are no direct constraints on the values of the choice variables. Second, it assumes that constraints cannot be satisfied as inequalities. Let us consider each of these assumptions in turn.

In many economic problems, variables are constrained to be non-negative. For example, consumers cannot consume negative quantities of goods, firms cannot produce negative outputs. In other problems there may be natural non-zero upper or lower bounds to the value of some variable, for example a shareholder can only hold a fraction between 0 and 1 of a company's shares. For the moment we concentrate on the case of non-negativity constraints.

In Fig. 2.12 of the previous section, used to motivate the discussion of Lagrange's method, we assumed that the tangency occurred at a point in the interior of the positive quadrant, with $x_1^* > 0$, $x_2^* > 0$. However, we could easily have drawn a diagram with the tangency at a point, say, with $x_1^* < 0$, $x_2^* > 0$. None of the ensuing discussion would have been affected. But clearly, if negative values of x_1 are ruled out, this could not be a solution to the problem, and we would have to think again. We do this with the help of an example.

Suppose $R(x)$ and $C(x)$ are a firm's revenue and cost functions respectively, where x is a vector of outputs (we have a multi-product firm). Then the firm's objective function is profit: $f(x) = R(x) - C(x)$. The unconstrained profit maximizing output vector is characterized by:

$$f_i(x^*) = R_i(x^*) - C_i(x^*) = 0 \quad [\text{H.1}]$$

which yields the familiar description of the profit maximizing output in terms of the equality of each output's marginal revenue ($R_i(x^*)$) with marginal cost ($C_i(x^*)$).

But is this problem really unconstrained? Given the inadmissibility of negative outputs we should impose the constraints:

$$x_i \geq 0 \quad i = 1, 2, \dots, n \quad [\text{H.2}]$$

The result of this is that the conditions in [H.1] are no longer necessary conditions for a maximum. To see this consider Fig. 2.14. In each part of the figure profit $f(x)$ is plotted holding all variables except the i th constant, at their optimal values. In (a) we have the case implicitly envisaged by the conditions in [H.1]. The i th output's optimal value $x_i^* > 0$

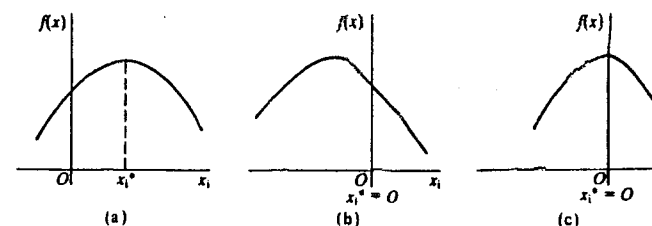


Fig. 2.14

$f(x)$ and $g^j(x)$ are all differentiable, this implies that x^* , λ^* are characterized by the following Kuhn-Tucker conditions:

$$\frac{\partial L}{\partial x_i} = f_i(x^*) + \sum_{j=1}^m \lambda_j^* g_i^j(x^*) \leq 0 \quad x_i^* \geq 0 \quad x_i^* \frac{\partial L}{\partial x_i} = 0 \quad i = 1, \dots, n \quad [\text{H.9}]$$

$$\frac{\partial L}{\partial \lambda_j} = g^j(x^*) \geq 0 \quad \lambda_j^* \geq 0 \quad \lambda_j^* g^j(x^*) = 0 \quad j = 1, \dots, m \quad [\text{H.10}]$$

The reason the partial derivatives $\partial L / \partial x_i$ must be non-positive at the optimum is because we are maximizing subject to the condition that $x_i \geq 0$, and so [H.9] reflects the same arguments that gave [H.4] earlier. Likewise, we require $\partial L / \partial \lambda_j \geq 0$ because we are minimizing subject to $\lambda_j \geq 0$ (recall [H.5]). The Kuhn-Tucker conditions then give us the required necessary conditions for the concave programming problem. We shall illustrate their use below, once we have examined in more depth the meaning and justification of this saddle point characterization of the optimum x^* .

The general idea of a saddle point is illustrated in Fig. 2.15. The function $f(x, y)$ is strictly concave in x , for each y , and strictly convex in y , for each x . A saddle point occurs at (x^*, y^*) : x^* maximizes $f(x, y^*)$, and y^* minimizes $f(x^*, y)$. The reason for the term 'saddle point' is obvious from the figure (a saddle point of a function need not look like that in Fig. 2.15, however – see Question 10, Exercise 2H).

Of course, any arbitrary function may not possess a saddle point – to do so would be quite a special property of the function. The Saddle Point Theorem given below is very important: it establishes that, if f and the g^j are concave functions, then there exists a vector of Lagrange multipliers λ^* which, together with the vector x^* which solves the concave programming problem [H.6], represent a saddle point of the Lagrange function, over the domain $x \geq 0$, $\lambda \geq 0$. So, given differentiability, x^* can then be characterized by the Kuhn-Tucker conditions [H.9] and [H.10].

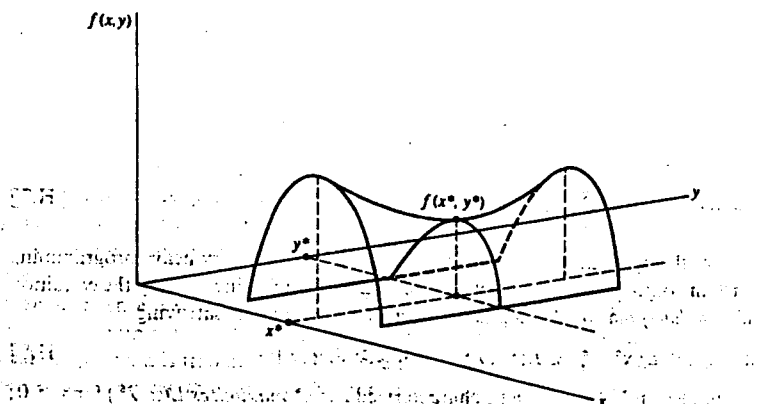


Fig. 2.15

The key result is given by the following

Fundamental Theorem of Concave Functions: Let $h^0(x)$, $h^1(x)$, ..., $h^m(x)$ be concave functions defined on a convex set X . If there is no $\bar{x} \in X$ such that $h^j(\bar{x}) > 0$, all $j = 0, 1, \dots, m$, then there exist numbers $p_j \geq 0$, not all zero, such that

$$\sum_{j=0}^m p_j h^j(x) \leq 0 \quad \text{for all } x \in X \quad [\text{H.11}]$$

In words, if there is no point in the domain at which all the functions simultaneously take on a positive value, then we can always find a set of non-negative weights such that the weighted sum of the functions is never positive, whatever the value of x . It is easy to see why we have to rule out the existence of a point like \bar{x} , since then obviously no such p_j could exist. The interesting thing is that fixed weights can be found that keep the weighted sum non-positive for all values of $x \in X$.

Figure 2.16 illustrates for three strictly concave functions. At x' , $h^1(x') \geq 0$ and $h^2(x') > 0$, but $h^0(x') < 0$, and the functions satisfy the conditions of the theorem at all other points also. $X = \{x | 0 \leq x \leq 1\}$. Then clearly choosing $p_0 = 1$, $p_1 = 1$, and $p_2 = 0$ would do the trick – at no x is $1 \cdot h^0(x) + 1 \cdot h^1(x) + 0 \cdot h^2(x) > 0$. (Explain why in this example we could never set $p_0 = 0$ if we want to satisfy [H.11]).

This theorem yields the central result on the solution to the concave programming problem [H.6] directly.

Saddle-point Theorem: if x^* is a solution to the concave programming problem [H.6], and Slater's condition S: there exists \bar{x} in X such that $g^j(\bar{x}) > 0$, $j = 1, \dots, m$ is satisfied, then there exist multipliers $\lambda_j^* \geq 0$, $j = 1, \dots, m$, not all zero, such that (x^*, λ^*) is a saddle point of the Lagrange function, i.e.

$$L(x, \lambda^*) \leq L(x^*, \lambda^*) \leq L(x^*, \lambda) \quad \text{for } x \geq 0, \lambda \geq 0$$

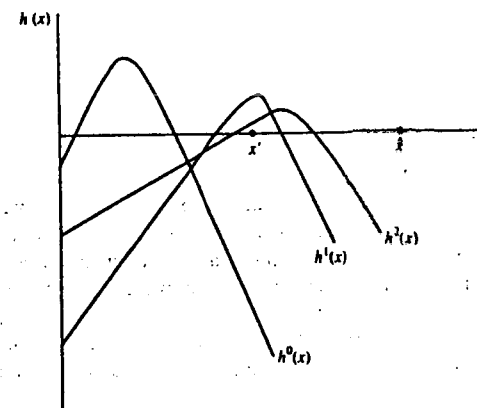


Fig. 2.16

Proof: since $f(x^*) \geq f(x)$ for all $x \geq 0$ such that $g^j(x) \geq 0$, then the system

$$g^j(x) > 0 \quad j = 1, \dots, m$$

$$f(x) - f(x^*) > 0 \quad [\text{H.12}]$$

has no solution for $x \geq 0$. The g^j and f are all concave and so the conditions of the previous theorem are satisfied with $X = R_+^n$, and so there exist numbers $p_0^*, p_1^*, \dots, p_m^* \geq 0$, not all zero, such that

$$p_0^*[f(x) - f(x^*)] + \sum_{j=1}^m p_j^* g^j(x) \leq 0 \quad \text{for all } x \in X \quad [\text{H.13}]$$

Since this must also be true for $x = x^*$, we have $\sum_{j=1}^m p_j^* g^j(x^*) \leq 0$. But since x^* is feasible, $p_j^* \geq 0$ and $g^j(x^*) \geq 0$ imply $\sum_{j=1}^m p_j^* g^j(x^*) \geq 0$, so these two weak inequalities together mean $\sum_{j=1}^m p_j^* g^j(x^*) = 0$. Thus we can rewrite [H.13] as

$$p_0^* f(x) + \sum_{j=1}^m p_j^* g^j(x) \leq p_0^* f(x^*) + \sum_{j=1}^m p_j^* g^j(x^*) \quad \text{all } x \in X \quad [\text{H.14}]$$

Note that $g^j(x^*) \geq 0$ and $p_j \geq 0$ implies $\sum_{j=1}^m p_j g^j(x^*) \geq 0$, so that we can re-write [H.14] as

$$p_0^* f(x) + \sum_{j=1}^m p_j^* g^j(x) \leq p_0^* f(x^*) + \sum_{j=1}^m p_j^* g^j(x^*) \leq p_0^* f(x^*) + \sum_{j=1}^m p_j g^j(x^*)$$

$$\text{all } x \in X, p_j \geq 0 \quad [\text{H.15}]$$

Now suppose $p_0^* = 0$. Then [H.13] implies $\sum_{j=1}^m p_j^* g^j(x) \leq 0$. In particular at $x = \bar{x}$ we have $\sum_{j=1}^m p_j^* g^j(\bar{x}) \leq 0$. But since each $p_j^* \geq 0$ and not all are zero, this must violate Slater's condition and we have a contradiction. Thus Slater's condition implies $p_0^* > 0$. So we can define $\lambda_j^* = p_j^*/p_0^*$, $\lambda_j = p_j/p_0^*$, $j = 1, \dots, m$, and rewrite [H.15] as

$$f(x) + \sum_{j=1}^m \lambda_j^* g^j(x) \leq f(x^*) + \sum_{j=1}^m \lambda_j^* g^j(x^*) \leq f(x^*) + \sum_{j=1}^m \lambda_j g^j(x^*)$$

$$\text{all } x \in X, \lambda_j \geq 0 \quad [\text{H.16}]$$

Then, recalling the definition of the Lagrange function $L(x, \lambda)$, we see that [H.16] is the saddle-point result.

Note that Slater's condition was not required to derive [H.15], but only when we wanted to move to the Lagrange function form [H.16]. Slater's condition can be interpreted as requiring that the feasible set possesses an interior point, and is designed to rule out the following kind of case, in which the Lagrange procedure breaks down. Suppose the problem is $\max x$ s.t. $-x^2 \geq 0$, $x \in R$. The Lagrange function is $x - \lambda x^2$, and the 'necessary conditions' are $1 - 2\lambda x \leq 0$, $-x^2 \geq 0$, but since the only x which satisfies the constraint is $x = 0$, this leads to the nonsense result $1 \leq 0$. The solution is obviously $x^* = 0$, but no $\lambda^* \geq 0$ exists such that

$$x - \lambda^* x^2 \leq x^* - \lambda^* x^{*2} \leq x^* - \lambda x^{*2}$$

If $\lambda^* = 0$, the first inequality is violated by choosing $x > 0$; if $\lambda^* > 0$, then that inequality is violated by choosing $x < 1/\lambda^*$. Thus the saddle-point condition does not characterize the solution of this concave programming problem. Slater's condition is clearly not satisfied (no x exists for which $-x^2 > 0$). Note, however, it is true that

$$p_0^* x - p_1^* x^2 \leq p_0^* x^* - p_1^* x^{*2} \leq p_0^* x^* - p_1 x^{*2}$$

for $p_0^* = 0$, $x^* = 0$, and any $p_1^* > 0$. In terms of our illustration in Fig. 2.16, Slater's condition is ensuring that it is the objective function that corresponds to the function $h^0(x)$ and not one of the constraints. In other words, it is ensuring that p_0^* cannot be set to zero when taking the weighted sum of the concave functions. Hence it is permissible to divide through by p_0 in deriving the Lagrange multipliers.

It is also possible to prove the converse of this theorem, namely that if (x^*, λ^*) is a saddle point of the Lagrange function then x^* solves the problem in [H.6], and this does not require concavity of the functions f, g^j . Taking the two theorems together, the Kuhn-Tucker conditions are both necessary and sufficient conditions for a solution x^* to problem [H.6] when the functions f, g^j are concave. It is also possible to extend these theorems to the case in which any of the functions are quasi-concave. This is particularly important because ordinal utility functions in economics cannot be restricted to be concave, but only quasi-concave. It is then reassuring to know that the Kuhn-Tucker conditions are directly applicable (the interested reader is referred to Takayama, 1985, Ch. 1).

To illustrate the use of the Kuhn-Tucker conditions we consider two problems, the first involving only non-negativity conditions, the second introducing also weak inequalities in the functional constraints. Consider the two-variable problem:

$$\max f(x_1, x_2) \text{ s.t. } a_1 x_1 + a_2 x_2 = b \quad x_1, x_2 \geq 0 \quad [\text{H.17}]$$

where f is taken to be strictly increasing and strictly quasi-concave. However, we assume that the contours of the objective function are everywhere steeper than the constraint line. Fig. 2.17 illustrates. The Lagrange function for the problem is:

$$L(x_1, x_2, \lambda) = f(x_1, x_2) - \lambda[a_1 x_1 + a_2 x_2 - b] \quad [\text{H.18}]$$

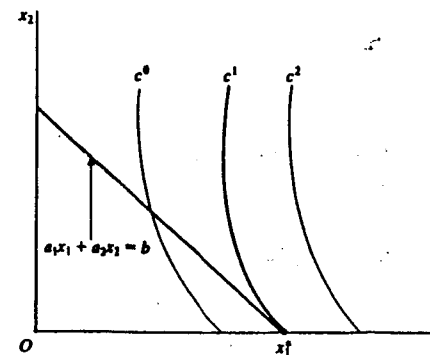


Fig. 2.17

and so the counterparts to condition [H.4] are:

$$L_1 = f_1 - \lambda^* a_1 \leq 0 \quad x_1^* \geq 0 \quad x_1^* \cdot [f_1 - \lambda^* a_1] = 0 \quad [\text{H.19}]$$

$$L_2 = f_2 - \lambda^* a_2 \leq 0 \quad x_2^* \geq 0 \quad x_2^* \cdot [f_2 - \lambda^* a_2] = 0 \quad [\text{H.20}]$$

$$L_3 = -(a_1 x_1^* + a_2 x_2^* - b) = 0 \quad [\text{H.21}]$$

Now suppose, as in the figure, that at the optimum, $x_1^* > 0$ and $x_2^* = 0$. From [H.19] we must have (explain why):

$$f_1 = \lambda^* a_1 \quad [\text{H.22}]$$

while from [H.20] we have:

$$f_2 \leq \lambda^* a_2 \quad [\text{H.23}]$$

Dividing each side of [H.23] into the corresponding side of [H.22] gives

$$f_1/f_2 \geq a_1/a_2 \quad [\text{H.24}]$$

which is simply the condition that, at the optimum, the contour of the objective function must be at least as steep as the constraint line. This is in fact all that can be said in characterizing an optimum when non-negativity constraints exist and one is binding at the optimum. Note that if at the optimum $x_2^* > 0$, then we have of course the necessary conditions in the form given in section G.

Turning now to the case of inequalities in the functional constraints, we can note first that in single-constraint problems the non-existence of bliss points makes this generalization unnecessary. Since in this case a solution will always lie on the boundary, we might as well express the constraint in equality form, as in the problem [H.17]. The generalization does, however, become important in problems of two or more constraints.

Consider the problem:

$$\begin{aligned} \max f(x_1, x_2) \quad \text{s.t.} \quad & a_1 x_1 + a_2 x_2 \leq b_1 \\ & c_1 x_1 + c_2 x_2 \leq b_2 \quad x_1, x_2 \geq 0 \end{aligned} \quad [\text{H.25}]$$

where f is concave and $f_1, f_2 > 0$. The problem is illustrated in Fig. 2.18. It is assumed that the constraints are such as to intersect in the positive quadrant. The feasible set is then the shaded area. Points α , β and γ correspond to possible types of solution for different assumptions about the contours of the objective function, and assuming the non-negativity constraints are non-binding at the optimum.

The Lagrange function is:

$$L(x_1, x_2, \lambda_1, \lambda_2) = f(x_1, x_2) - \lambda_1 [a_1 x_1 + a_2 x_2 - b_1] - \lambda_2 [c_1 x_1 + c_2 x_2 - b_2]$$

and the Kuhn-Tucker conditions are:

$$f_1 - \lambda_1^* a_1 - \lambda_2^* c_1 \leq 0 \quad x_1^* \geq 0 \quad x_1^* [f_1 - \lambda_1^* a_1 - \lambda_2^* c_1] = 0 \quad [\text{H.27}]$$

$$f_2 - \lambda_1^* a_2 - \lambda_2^* c_2 \leq 0 \quad x_2^* \geq 0 \quad x_2^* [f_2 - \lambda_1^* a_2 - \lambda_2^* c_2] = 0 \quad [\text{H.28}]$$

$$a_1 x_1^* + a_2 x_2^* - b_1 \leq 0 \quad \lambda_1^* \geq 0 \quad \lambda_1^* [a_1 x_1^* + a_2 x_2^* - b_1] = 0 \quad [\text{H.29}]$$

$$c_1 x_1^* + c_2 x_2^* - b_2 \leq 0 \quad \lambda_2^* \geq 0 \quad \lambda_2^* [c_1 x_1^* + c_2 x_2^* - b_2] = 0 \quad [\text{H.30}]$$

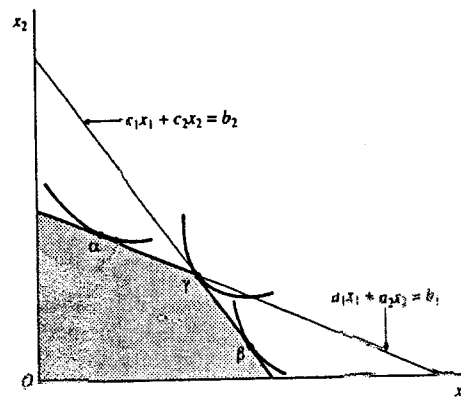


Fig. 2.18

Let us now establish the connection between the solution possibilities in the figure and the conditions in [H.27]–[H.30]. In case of a solution at α , we have that the b_2 constraint is non-binding. It follows therefore that $\lambda_2^* = 0$ (intuitively, the shadow price of a non-binding constraint is zero). Therefore the terms involving λ_2^* in [H.27] and [H.28] drop out and we have the conditions for

$$\text{case } \alpha \quad f_1 - \lambda_1^* a_1 = 0 \quad f_2 - \lambda_1^* a_2 = 0 \quad a_1 x_1^* + a_2 x_2^* = b_1 \quad [\text{H.31}]$$

These are the standard conditions resulting from applying the Lagrange procedure to the appropriate single-constraint problem, i.e. that with only the b_1 constraint expressed as an equality.

Consider now case β . Here, the b_1 constraint is non-binding. Therefore small shifts in it have no effect on the optimum and $\lambda_1^* = 0$. In that case we have the conditions for:

$$\text{case } \beta \quad f_1 - \lambda_2^* c_1 = 0 \quad f_2 - \lambda_2^* c_2 = 0 \quad c_1 x_1^* + c_2 x_2^* = b_2 \quad [\text{H.32}]$$

Again, therefore, we obtain the appropriate conditions from applying the Lagrange procedure to the appropriate single-constraint problem.

In the third case γ both constraints are binding at the optimum, and are satisfied as equalities (but see Question 7*, Exercise 2H). Strictly speaking, conditions [H.29] and [H.30], the solution of which determines the point (x_1^*, x_2^*) , are in the present case sufficient to solve the problem. From [H.27] and [H.28] we see that:

$$f_1/f_2 = \frac{\lambda_1^* a_1 + \lambda_2^* c_1}{\lambda_1^* a_2 + \lambda_2^* c_2} \quad [\text{H.33}]$$

i.e. at the optimum the slope of the contour of the objective function is *not equal* to the slope of either constraint, but rather lies between these slopes (see Question 9*, Exercise 2H).

There are two remaining solution possibilities, namely those in which one of the variables is zero at the optimum, but considering these would add nothing to the discussion of the previous example (confirm).

From the Implicit Function Theorem, introduced in section G, we know that if the determinant of partial derivatives

$$D = \begin{bmatrix} f_1^1 & \dots & f_n^1 \\ \dots & \dots & \dots \\ f_1^n & \dots & f_n^n \end{bmatrix}$$

is non-zero, then there exist n functions $h^i(\alpha_1, \dots, \alpha_m)$, such that $x_i = h^i(\alpha_1, \dots, \alpha_m)$, $i = 1, \dots, n$. We think of the equations in [1.1] as determining the equilibrium values of the x_i , and the theorem then tells us that these equilibrium values can be regarded as functions of the parameters α_j , $j = 1, \dots, m$. The problem of comparative statics is to say as much as we can about the derivatives $h_j^i = \partial x_i / \partial \alpha_j$, since these express the effects of a change in a parameter on the equilibrium value of a variable.

We proceed as follows. Totally differentiating through the system [I.1] gives

$$f_1^1 dx_1 + \cdots + f_n^1 dx_n + f_{n+1}^1 d\alpha_1 + \cdots + f_{n+m}^1 d\alpha_m = 0$$

..... [I.2]

$$f_1^n dx_1 + \dots + f_n^n dx_n + f_{n+1}^n d\alpha_1 + \dots + f_{n+m}^n d\alpha_m = 0$$

The total differentials of the f^i functions are set to zero because we constrain the equilibrium conditions to continue to hold when we make some specified change dx_j in a parameter, where these dx_j are given (small) numbers. That is, the differentials dx_i must be such as to keep the equilibrium conditions satisfied when the α_j change. Moreover, all the partial derivatives in [1.2] are evaluated at the initial equilibrium values of the x_i and given values α_j , $i = 1, \dots, n$, $j = 1, \dots, m$. This means that those partial derivatives are just numbers, and so [1.2] can be regarded as a *system of linear equations*, of the form

$$\begin{bmatrix} f_1^1 & \dots & f_n^1 \\ \dots & \dots & \dots \\ f_1^n & \dots & f_n^n \end{bmatrix} \begin{bmatrix} dx_1 \\ \vdots \\ dx_n \end{bmatrix} = \begin{bmatrix} -(f_{n+1}^1 d\alpha_1 + \dots + f_{n+m}^1 d\alpha_m) \\ \dots \\ -(f_{n+1}^n d\alpha_1 + \dots + f_{n+m}^n d\alpha_m) \end{bmatrix} \quad [1.3]$$

Then, we wish to solve for the unknowns, the dx_i , in terms of the given changes. Suppose that only one parameter has changed, i.e. $dx_j \neq 0$, $dx_k = 0$, $k \neq j$. Then, from Cramer's Rule, the solution for any dx_i is given by

$$dx_i = \frac{D_{ij}}{D} dx_j \quad \text{or,} \quad \frac{\partial x_i}{\partial \alpha_i} = \frac{D_{ij}}{D} \quad [1.4]$$

where D_{ij} is the determinant formed by replacing the i th column of the determinant D (defined above and assumed non-zero) by the column vector $[-f_{n+j}^1, \dots, -f_{n+j}^n]$.

At the level of generality of most economic models, all we have are general restrictions on the signs of the partial derivatives of the f^i functions. The most that can be expected is that we can deduce from [I.4] the sign of the comparative statics effect $\partial x_i / \partial \alpha_j$. Even this much is often not possible: we may have to develop a taxonomy of cases in which the comparative statics effect is positive or negative.

To illustrate, suppose we have a problem of the form

$$\max_{x_1, x_2} u(x_1, x_2) \text{ s.t. } \alpha_1 x_1 + \alpha_2 x_2 = \alpha_3$$

Then the first-order conditions are

$$\begin{aligned} u_1(x_1^*, x_2^*) - \lambda^* \alpha_1 &= 0 \\ u_2(x_1^*, x_2^*) - \lambda^* \alpha_2 &= 0 \\ -\alpha_1 x_1^* - \alpha_2 x_2^* + \alpha_3 &= 0 \end{aligned} \quad [1.5]$$

where asterisks denote equilibrium values. Here, the counterparts of the f^i functions in [1.1] are the partial derivatives of the Lagrange function, $\partial L/\partial x_i$, $i = 1, 2$, and $\partial L/\partial \lambda$, and the endogenous variables are x_1 , x_2 , and λ . Then, carrying out the total differentiation gives the linear system

$$\begin{bmatrix} u_{11} & u_{12} & -\alpha_1 \\ u_{21} & u_{22} & -\alpha_2 \\ -\alpha_1 & -\alpha_2 & 0 \end{bmatrix} \begin{bmatrix} dx_1 \\ dx_2 \\ d\lambda \end{bmatrix} = \begin{bmatrix} \lambda^* dx_1 \\ \lambda^* dx_2 \\ x_1^* dx_1 + x_2^* dx_2 - d\alpha_3 \end{bmatrix} \quad [1.6]$$

Solving for x_1 , considering first only $dx_1 \neq 0$, then $dx_3 \neq 0$, gives

$$\frac{\partial x_1}{\partial \alpha_1} = \frac{-\lambda^* \alpha_2^2 + x_1^* (\alpha_1 u_{22} - \alpha_2 u_{12})}{D}, \quad \frac{\partial x_1}{\partial \alpha_3} = \frac{-(\alpha_1 u_{21} - \alpha_2 u_{12})}{D} \quad [1.7]$$

where D is the determinant of the matrix on the left-hand side of [16] and must of course be assumed non-zero.

In order to put signs to the terms in [1.7] we need to know the signs of D , α_1 , α_2 , u_{12} , u_{22} and λ^* . The economics underlying the formulation of the problem may help us in this, but note that the expressions in [1.7] involve sums and differences of terms and so we may well not be able to obtain unique signs for the partial derivatives (see Question 1, Exercise 21). For our present purposes, however, it is enough to note that the sign of D must follow from the second order conditions of the problem, and so, having motivated the study of second order conditions, we now turn to this.

Second order conditions for unconstrained maximization

In the problem of the *unconstrained* maximization of a function of a single variable, $f(x)$, the role of a second order condition is to determine whether a point x^* which yields a stationary value of the function, $f'(x^*) = 0$, does in fact maximize the function. If the second-order condition $f''(x^*) < 0$ holds, then the first- and second-order conditions taken together are *sufficient* for x^* to be a locally maximizing point. They are *not necessary*, since we may have a function for which a point \hat{x} yields a maximum but $f''(\hat{x}) = 0$, and we would have to examine higher order derivatives to establish its optimality (we meet an example below). Such cases are excluded when we do comparative statics, for reasons which will soon become clear, and so we focus on the sufficient second-order condition where the strict inequality holds.

As an example, suppose a firm has a revenue function $\bar{p}x$, where $\bar{p} = 500$ is a constant price and x is output, and a total cost function $C = 750x - 30x^2 + 0.5x^3$. This cubic cost function gives the 'usual' U-shaped average and marginal curves. The firm chooses x

to maximize profit $\pi = \bar{p}x - C = (500 - 750)x + 30x^2 - 0.5x^3$, giving the first order condition

$$-250 + 60x - 1.5x^2 = 0$$

and yielding solution values $x^0 = 4.72$, $x^1 = 35.27$. The second derivative of the profit function is $60 - 3x$, and so, clearly, only $x^1 = 35.27$ satisfies the second-order condition and gives a local maximum (for further discussion and illustration of this problem see section 9A).

Comparative statics: single choice variable

To illustrate the use of the second-order conditions in comparative static analysis, suppose a decision-maker wishes to maximize a function $f(x, \alpha)$, where the single-choice variable is the scalar x and α is a parameter. Let x^* be a solution to this problem, so that x^* satisfies the first-order condition $f_x(x^*, \alpha) = 0$. When α alters the decision-maker will change x^* so that the first-order condition continues to hold. Thus it must be true that the total change in f_x must be zero or

$$df_x(x^*, \alpha) = f_{xx}(x^*, \alpha)dx^* + f_{x\alpha}(x^*, \alpha)d\alpha = 0$$

Rearranging this gives

$$\frac{dx^*}{d\alpha} = -\frac{f_{x\alpha}}{f_{xx}} \quad [I.8]$$

Note that to obtain this result we must assume $f_{xx} \neq 0$. If we make this assumption then at the optimum $f_{xx} < 0$ and so the sign of $dx^*/d\alpha$ in [I.8] is the same as the sign of $f_{x\alpha}$.

Although this appears to be very simple we will frequently find that it is possible to formulate interesting economic problems as single-variable decision problems, and we will make repeated use of [I.8] in later chapters. Figure 2.19 provides some intuition for the result. The decision-maker initially faces a decision problem in which $\alpha = \alpha^0$ and maximizes $f(x, \alpha^0)$ by setting $x = x^*(\alpha^0)$ where $f_x(x, \alpha^0)$, the marginal value of x , is zero. We assume that increases in the parameter α increase the marginal value of x and thus, when α increases from α^0 to α^1 , the optimal x increases to $x^*(\alpha^1)$. We see that if the parameter change increases the marginal value of the decision variable the decision-maker will increase x , since $f_x(x^*(\alpha^0), \alpha^1)$, the marginal value of x at its initial optimal level, is now positive.

The cases we have to exclude to apply this procedure are not unduly strange. Take for example $f(x; \alpha) = -\alpha x^4$. This function is maximized at $x^* = 0$, but at this point its second derivative $-12\alpha x^2 = 0$, and so $-f_{x\alpha}/f_{xx}$ in [I.8] would be undefined. Thus in the general analysis we have to exclude this kind of case (but it is not at all difficult to give the effect of a change in α on x^* - what is it?).

Second order conditions: n choice variables*

We now extend the discussion to the case of the maximization of a function of an n -vector of variables, $f(x)$. Suppose this function has a maximum at x^* , so that $f_i(x^*) = 0$, $i = 1, \dots, n$. Taking a Taylor series expansion of the function around this point gives, to

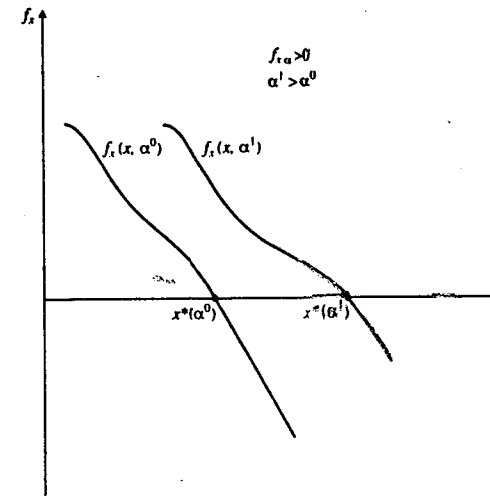


Fig. 2.19

the second order

$$f(x^* + h) = f(x^*) + h\nabla f + (1/2)hFh \quad [I.9]$$

where ∇f is the vector of first order partials and is therefore zero, h is an n -vector (h_1, \dots, h_n) of small numbers, and F is the Hessian matrix $[f_{ij}]$, $i, j = 1, \dots, n$, i.e. the $n \times n$ symmetric matrix of second-order partials, evaluated at x^* . It is then sufficient for x^* to yield a local maximum that

$$hFh = (h_1, \dots, h_n) \begin{bmatrix} f_{11} & \dots & f_{1n} \\ \dots & \dots & \dots \\ f_{n1} & \dots & f_{nn} \end{bmatrix} \begin{bmatrix} h_1 \\ \vdots \\ h_n \end{bmatrix} = \sum_{i=1}^n \sum_{j=1}^n f_{ij}h_ih_j < 0 \quad [I.10]$$

since then, from [I.9], $f(x^* + h) < f(x^*)$ for any small (non-zero) vector h . Again, the conditions $\nabla f = 0$, $hFh < 0$ are not necessary for an optimum because a maximum could occur at a point at which $hFh = 0$. However, for purposes of comparative statics we have to rule out cases in which this happens, and so we focus on the sufficient second-order condition $hFh < 0$.

The expression in [I.10] is a quadratic form, and condition [I.10] is the condition that this quadratic form be negative definite. From the theory of quadratic forms we have the following sufficient condition for this. A principal minor of order $k = 1, \dots, n$, of a determinant is the sub-determinant formed by deleting the last $n - k$ rows and columns. Thus the principal minors of the $n \times n$ determinant $|F|$ are

$$|f_{11}|, \begin{vmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \end{vmatrix}, \dots, \begin{vmatrix} f_{11} & \dots & f_{1k} \\ \dots & \dots & \dots \\ f_{k1} & \dots & f_{kk} \end{vmatrix}, \dots, \begin{vmatrix} f_{11} & \dots & f_{1n} \\ \dots & \dots & \dots \\ f_{n1} & \dots & f_{nn} \end{vmatrix} \quad [I.11]$$

Then, the quadratic form in [I.10] is negative definite if the k th principal minor of $|F|$ has the sign $(-1)^k$, $k = 1, \dots, n$.

This requires that the principal minors alternate in sign:

$$f_{11} < 0; \quad f_{11}f_{22} - f_{21}f_{12} > 0; \dots \quad [\text{I.12}]$$

[I.12] then gives the second-order sufficient condition for a maximum in the n -variable case.

This suggests that if we seek to maximize $f(x)$, we first solve the n equations $f_i(x^*) = 0$ for x^* , then evaluate the signs of the n principal minors in [I.12] by plugging x^* into the partial derivatives $f_{ij}(x)$, to check that we do in fact have a local maximum.

In economics, we often take a different approach. We appeal to some aspect of the economic nature of the problem (diminishing marginal productivity; diminishing returns to scale; diminishing marginal utility, and so on) to make a *global* sufficiency assumption (the conditions [I.12] are *local* since they were derived from an expansion in a neighbourhood of x^*). The assumption is that the objective function $f(x)$ is strictly concave over its domain. The stationary value of a strictly concave function must be a global maximum, and so the point x^* at which $f_i(x^*) = 0$, $i = 1, \dots, n$, must be also a local maximum. Diagrammatically, the graph of the function in three dimensions is like a hill or dome, the tangent plane to the peak of the hill is horizontal, and no other tangent plane to the hill has this property. Thus, if we make this global sufficiency assumption, which has the attraction that we then know that any optimum is a true global optimum, we do not need *local* second-order conditions.

Comparative static analysis: n choice variables*

The local conditions are useful primarily in comparative statics analysis which involves local changes in the parameter values. If a function $f(x)$ is concave it can be shown that at any point in its domain, including its maximum, we have

$$hFh = \sum_{i=1}^n \sum_{j=1}^n f_{ij}h_ih_j \leq 0 \quad [\text{I.13}]$$

That is, the quadratic form involving the Hessian matrix F at any point of the domain of the function is *negative semidefinite*. Unfortunately, it is not true in general that for a *strictly* concave function, [I.13] is satisfied as a strict inequality, i.e. that hFh is negative definite. However, apart from cases in which $hFh = 0$ at the optimal point x^* the global sufficiency assumption of strict concavity implies that the local sufficiency condition [I.12] is satisfied and F has the required properties. We now illustrate this with an example.

A decision-taker wants to maximize the function $f(x_1, x_2, \alpha)$, which is assumed strictly concave in x_1 and x_2 . The first-order conditions are

$$\begin{aligned} f_1(x_1^*, x_2^*, \alpha) &= 0 \\ f_2(x_1^*, x_2^*, \alpha) &= 0 \end{aligned} \quad [\text{I.14}]$$

Differentiating totally through these two conditions gives the system

$$\begin{bmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \end{bmatrix} \begin{bmatrix} dx_1^* \\ dx_2^* \end{bmatrix} = \begin{bmatrix} -f_{1\alpha}d\alpha \\ -f_{2\alpha}d\alpha \end{bmatrix} \quad [\text{I.15}]$$

Using Cramer's Rule to solve, say, for dx_1^* gives

$$\frac{dx_1^*}{d\alpha} = \frac{f_{12}f_{2\alpha} - f_{22}f_{1\alpha}}{|F|} \quad [\text{I.16}]$$

where $|F|$ is the Hessian determinant and must be assumed non-zero at the optimal point. This assumption, together with the strict concavity of f , implies that hFh is negative definite at (x_1^*, x_2^*) , and this allows us to sign the denominator in [I.16]. From the condition on principal minors we know that if hFh is negative definite

$$f_{11} < 0, \quad f_{11}f_{22} - f_{21}f_{12} > 0$$

and so the denominator in [I.16] is positive. We can then proceed to discuss the sign of the numerator (which may not be uniquely defined).

For a case in which this procedure breaks down, take the case in which $f = -\alpha(x_1^* + x_2^*)$ which is strictly concave and has a maximum at $x_1^* = x_2^* = 0$. The intuition is just as in the one-variable case.

We can briefly state the second-order conditions for the case in which we minimize a function $f(x)$, where x has $n \geq 2$ components. Proceeding from the derivation of [I.9] above, we have that sufficient conditions for x^* to yield a minimum of the function are that $\nabla f = 0$ and the quadratic form $hFh > 0$ at x^* . A sufficient condition for this quadratic form to be positive definite is that *all* the principal minors of $|F|$ be strictly positive. This is the local second-order sufficient condition for a minimum. A global sufficient condition is that f be strictly convex over its domain. Excluding the cases in which $hFh = 0$ at x^* (for example the case $f(x) = x_1^* + x_2^*$) this will imply that $hFh > 0$ at the optimum and that fact can be used directly in the comparative statics analysis of the minimization problem.

Second-order conditions for constrained maximization*

The formal statement of second-order conditions for the constrained maximization problem can be quite complex, but the essential ideas can be brought out simply if we first consider the two-variable one-constraint case. Thus suppose the problem is

$$\max_{x_1, x_2} f(x_1, x_2) \text{ s.t. } g(x_1, x_2) = 0 \quad [\text{I.17}]$$

and refer to Fig. 2.20. In each case in the figure a local solution to the problem is shown as the point of tangency, x^* , between a contour of f and the constraint contour. Since x^* is a *constrained* local maximum, it must not be possible to reach a higher contour of f (we assume $f_i > 0$, $i = 1, 2$) by moving along the constraint contour – the only feasible variations in the x_i . This is the case in each part of the figure. In (a) we have that f is strictly quasi-concave and g is either strictly quasi-convex (\bar{g}) or linear (\bar{g}). In (b) both functions are strictly quasi-concave, but around the optimum x^* , the contour of f is 'more curved' than that of g . In (c) both functions are strictly quasi-convex but around x^* the f -contour is 'less curved' than that of g . In each case, small movements along the g -contour away from x^* must reduce the value of the objective function. This, geometrically, is the essence of the second-order condition. We can say that if x_1 increases (decreases) from x_1^* , then the slope of the g -contour must become greater (smaller) in absolute value, or

smaller (greater) algebraically, than the slope of the f -contour. Since the slopes are equal at x^* , we can write this as the condition.

$$\frac{d}{dx_1} \left[\frac{dx_2}{dx_1} \right]_f - \frac{dx_2}{dx_1} \Big|_g > 0 \quad \text{at } x = x^* \quad [I.18]$$

where $dx_2/dx_1|_f$ denotes the slope of the f -contour, and $dx_2/dx_1|_g$ that of the g -contour. Increasing x_1 from x_1^* increases $dx_2/dx_1|_f$ relative to $dx_2/dx_1|_g$, since the former becomes 'less negative' than the latter.

We can derive [I.18] more formally as follows. Assume $g_2 \neq 0$. Then we can solve $g(x_1, x_2) = 0$ for x_2 as a function of x_1 , $x_2 = h(x_1)$, where $h' = dx_2/dx_1 = -g_1/g_2$. Substituting into f gives $f(x_1, h(x_1)) \equiv \phi(x_1)$. The value x_1^* which maximizes ϕ satisfies $\phi'(x_1^*) = 0$, $\phi''(x_1^*) < 0$ (again excluding cases where $\phi''(x_1^*) = 0$). But this second order derivative is

$$\phi''(x_1^*) = \frac{d}{dx_1} [f_1 + f_2 h'(x_1^*)] = f_{11} + f_{12} h' + h'(f_{21} + f_{22} h') + f_2 h'' < 0 \quad [I.19]$$

At the optimal point, $-g_1/g_2 = -f_1/f_2 = h'$, and substituting this into [I.19] and rearranging gives

$$\frac{1}{f_2^2} \{ f_{11} f_2^2 - 2 f_{12} f_1 f_2 + f_1^2 f_{22} \} + \frac{d^2 x_2}{dx_1^2} \Big|_g < 0 \quad [I.20]$$

where we have also used $h'' = d^2 x_2/dx_1^2$. But the first term in [I.20] is $-d^2 x_2/dx_1^2|_f$ (recall Question 7, Exercise 2B). Thus we have established [I.18].

Note that, as Fig. 2.20 shows, [I.18] may hold even though f is not quasi-concave or g is not quasi-convex. Therefore the local sufficiency condition [I.18] is less restrictive than the global condition that f be strictly quasi-concave and g be quasi-convex over their entire domains. However, we do usually assume this global condition, in order to ensure local maxima are also global maxima (in cases (b) and (c) this may well not be the case) and so we at the same time ensure satisfaction of the local second-order condition [I.18].

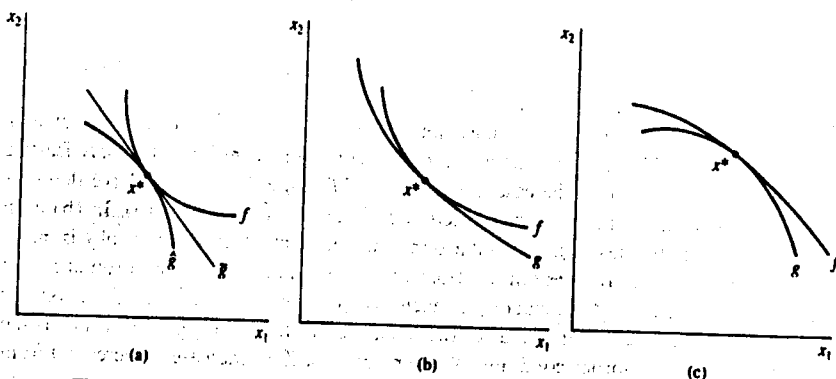


Fig. 2.20

Though perfectly adequate for the two-variable case, condition [I.18] does not generalize readily to $n > 2$ variables, and it is usual to express the second-order conditions in terms of the signs of the principal minors of a particular determinant analogously to the unconstrained case. This is also the more useful form for comparative statics analysis. A full treatment of these determinantal conditions cannot be given here. We shall simply state the conditions for the general case, and then show that in the two-variable one-constraint case the condition is equivalent to [I.18].

Thus suppose we have the problem of maximizing a function of an n -vector of variables $f(x)$, subject to $m \leq n$ equality constraints $g^k(x) = 0$, $k = 1, \dots, m$. We form the Lagrange function as before: $L = f(x) - \sum_{k=1}^m \lambda_k g^k(x)$. The first-order conditions are then

$$\begin{aligned} L_i &= f_i(x^*) - \sum_{k=1}^m \lambda_k^* g_i^k(x^*) = 0 & i = 1, \dots, n \\ L_k &= -g^k(x^*) = 0 & k = 1, \dots, m \end{aligned} \quad [I.21]$$

at the optimal point x^* . $L_{ij} = f_{ij} - \sum_{k=1}^m \lambda_k^* g_{ij}^k$ denotes the second-order partial of L . We now define the $(m+n) \times (m+n)$ bordered Hessian matrix

$$H = \begin{bmatrix} L_{11} & \dots & L_{1n} & -g_1^1 & \dots & -g_1^m \\ \dots & \dots & \dots & \dots & \dots & \dots \\ L_{n1} & \dots & L_{nn} & -g_n^1 & \dots & -g_n^m \\ -g_1^1 & \dots & -g_1^n & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ -g_m^1 & \dots & -g_m^n & 0 & \dots & 0 \end{bmatrix} = \begin{bmatrix} L & G \\ G' & O \end{bmatrix} \quad [I.22]$$

H can be thought of as consisting of four sub-matrices: the $n \times n$ symmetric matrix of second order partials of the Lagrange function; the $n \times m$ matrix G of partials $L_{ki} = -g_i^k$; its transpose, G' ; and an $m \times m$ matrix of zeroes. The k th column of G , and corresponding row of G' , consists of the (negatives of the) n first derivatives of the constraint function g^k , $k = 1, \dots, m$, which are also the second order cross partials L_{ki} . We think of G , G' and O as forming a 'border' of the Hessian of the Lagrange function, L , hence the term bordered Hessian.

We now define a bordered principal minor of the Hessian determinant $|H|$ as a principal minor which has at least $m+1$ of the first rows and columns from L , and a border consisting of the appropriate rows and columns of partials $-g_i^k$, and of zeroes. Thus, let

$$|H_{m+1}| \equiv \begin{vmatrix} L_{11} & \dots & L_{1,m+1} & -g_1^1 & \dots & -g_1^m \\ \dots & \dots & \dots & \dots & \dots & \dots \\ L_{m+1,1} & \dots & L_{m+1,m+1} & -g_{m+1}^1 & \dots & -g_{m+1}^m \\ -g_1^1 & \dots & -g_{m+1}^1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ -g_1^m & \dots & -g_{m+1}^m & 0 & \dots & 0 \end{vmatrix} \quad [I.23]$$

$|H_{m+2}|$ is then constructed by adding the $m+2$ th row and column of L and adjusting the border appropriately, and so on. Thus we can define the sequence of bordered principal

minors

$$|H_{m+1}|, |H_{m+2}|, \dots, |H_{m+j}|, \dots, |H_{m+(n-m)}| \equiv |H|$$

The sufficient second-order conditions can then be given as follows: x^* yields a maximum in the problem with n variables and m ($< n$) constraints, if the bordered principal minor $|H_{m+j}|$ has the sign $(-1)^{m+j}$, for $j = 1, \dots, n-m$.

The reason we ignore principal minors with less than m rows and columns from L is that they are necessarily zero. The principal minor with exactly m rows and columns from L always has a sign which is independent of the terms L_{ij} .

Comparative statics analysis: constrained maximization*

As an illustration of the use of these conditions and their application in comparative statics, consider the problem with $n = 3$ and $m = 2$.

$$\max f(x_1, x_2, x_3) \text{ s.t. } g^1(x_1, x_2, x_3, \alpha) = 0; \quad g^2(x_1, x_2, x_3) = 0$$

where α is a parameter in the first constraint. The first order conditions are

$$\begin{aligned} L_i &= f_i - \lambda_1^* g_i^1 - \lambda_2^* g_i^2 = 0 & i = 1, 2, 3 \\ L_k &= -g^k = 0 & k = 1, 2 \end{aligned} \quad [I.24]$$

The (sufficient) second-order condition is then

$$\begin{vmatrix} L_{11} & L_{12} & L_{13} & -g_1^1 & -g_1^2 \\ L_{21} & L_{22} & L_{23} & -g_2^1 & -g_2^2 \\ L_{31} & L_{32} & L_{33} & -g_3^1 & -g_3^2 \\ -g_1^1 & -g_2^1 & -g_3^1 & 0 & 0 \\ -g_1^2 & -g_2^2 & -g_3^2 & 0 & 0 \end{vmatrix} < 0 \quad [I.25]$$

since $(-1)^{m+1} < 0$ for $m = 2$. Note that as we asserted earlier, the principal minor with 1 row and column from L is zero and the principal minor with $2 = m$ rows and columns from L does not depend on the L_{ij} terms:

$$\begin{vmatrix} L_{11} & -g_1^1 & -g_1^2 \\ -g_1^1 & 0 & 0 \\ -g_1^2 & 0 & 0 \end{vmatrix} = 0, \text{ and } \begin{vmatrix} L_{11} & L_{12} & -g_1^1 & -g_1^2 \\ L_{21} & L_{22} & -g_2^1 & -g_2^2 \\ -g_1^1 & -g_2^1 & 0 & 0 \\ -g_1^2 & -g_2^2 & 0 & 0 \end{vmatrix} = (g_1^1 g_2^2 - g_1^2 g_2^1)^2 > 0 \quad [I.26]$$

In this three-variable case $|H_{m+1}| = |H|$. To carry out the comparative statics, differentiate through the first-order conditions to obtain the linear system

$$\begin{bmatrix} L_{11} & L_{12} & L_{13} & -g_1^1 & -g_1^2 \\ L_{21} & L_{22} & L_{23} & -g_2^1 & -g_2^2 \\ L_{31} & L_{32} & L_{33} & -g_3^1 & -g_3^2 \\ -g_1^1 & -g_2^1 & -g_3^1 & 0 & 0 \\ -g_1^2 & -g_2^2 & -g_3^2 & 0 & 0 \end{bmatrix} \begin{bmatrix} dx_1 \\ dx_2 \\ dx_3 \\ d\lambda_1 \\ d\lambda_2 \end{bmatrix} = \begin{bmatrix} \lambda_1^* g_{1\alpha}^1 da \\ \lambda_1^* g_{1\alpha}^2 da \\ \lambda_1^* g_{2\alpha}^1 da \\ \lambda_1^* g_{2\alpha}^2 da \\ 0 \end{bmatrix} \quad [I.27]$$

which can be solved for dx_i/da if the left-hand matrix has non-zero determinant. But note that this determinant is precisely the bordered Hessian $|H|$, and the second-order conditions tell us that $|H| < 0$. Thus, using Cramer's Rule gives

$$\frac{dx_i}{da} = \frac{|H_i|}{|H|} \quad [I.28]$$

where $|H_i|$ is the determinant formed by replacing the i th column of $|H|$ by the column of coefficients of da in the vector on the right-hand side of [I.27]. Then, in [I.28] the sign of dx_i/da is the opposite of that of $|H_i|$ (which may well not be unambiguously determined, however).

We can see then from this example that the relevance of the second-order conditions for comparative statics analysis arises from the fact that the bordered Hessian H is always the matrix of second-order partials in the linear system we need to solve to obtain the comparative statics results.

To see the relationship between the second-order condition in determinant form and the curvature conditions illustrated in Fig. 2.20 and derived in [I.18], let us take the case in which $n = 2$, $m = 1$, as in [I.17]. The condition [I.25] then becomes

$$\begin{vmatrix} L_{11} & L_{12} & -g_1 \\ L_{21} & L_{22} & -g_2 \\ -g_1 & -g_2 & 0 \end{vmatrix} > 0 \quad [I.29]$$

since $(-1)^{m+1} > 0$ for $m = 1$. Expanding this determinant gives

$$-(g_1^2 L_{11} - 2g_1 g_2 L_{12} + g_2^2 L_{22}) > 0 \quad [I.30]$$

where we have used $L_{12} = L_{21}$. Since $L_{ij} = f_{ij} - \lambda g_{ij}$, $i, j = 1, 2$, by substituting these into [I.30] and rearranging we have

$$\frac{1}{\lambda} \{f_{11} g_2^2 - 2g_1 g_2 f_{12} + f_{22} g_1^2\} - \{g_{11} g_2^2 - 2g_1 g_2 g_{12} + g_{22} g_1^2\} < 0 \quad [I.31]$$

Then, using $g_i = f_i/\lambda$, $i = 1, 2$, and $\lambda^3 = (g_2/f_2)^3$ gives

$$\frac{1}{f_2^3} \{f_{11} f_2^2 - 2f_1 f_2 f_{12} + f_{22} f_1^2\} - \frac{1}{g_2^3} \{g_{11} g_2^2 - 2g_1 g_2 g_{12} + g_{22} g_1^2\} < 0 \quad [I.32]$$

implying

$$\left. \frac{d^2 x_2}{dx_1^2} \right|_f < \left. \frac{d^2 x_2}{dx_1^2} \right|_g \quad [I.33]$$

as in [I.18]. Thus, the determinantal condition is equivalent to the condition on the relative curvature of the f - and g -contours.

The concave programming case

In the problem with non-negativity conditions and weak inequalities in the constraints, we cannot simply differentiate through the Kuhn-Tucker conditions to carry out

comparative statics, because of the presence of inequalities in those conditions. Also, *global* sufficiency conditions are used. If the objective and constraint functions are all concave (the constraints are expressed as $g^j(x) \geq 0$) then the Kuhn-Tucker conditions are both necessary and sufficient for x^* to be a global maximum. If it is desired to carry out the conventional kind of comparative statics analysis, we exclude from the Kuhn-Tucker conditions those variables which are zero at the optimum, and those constraints that are non-binding there, so that in effect we end up with the conditions corresponding to a classical maximization problem. We can then apply the standard comparative statics procedure to these conditions. This involves the assumptions that for small changes, zero variables and non-binding constraints do not change their status.

Minimization problems

The problem of minimizing a function $f(x)$ where x is an n -vector, subject to $m < n$ constraints $g^k(x) = 0$, $k = 1, \dots, m$, has the same first-order conditions as the maximization problem. We could analyse the problem of second-order conditions for the case $n = 2$, $m = 1$, as we did for maximization in Fig. 2.19, and derive an analogous condition to [I.18]. This is left as an exercise. Here we simply state the sufficient local second order conditions in determinant form, and note that they play exactly the same role in the comparative statics analysis of minimization problems as was the case for maximization. These conditions are again framed in terms of the bordered Hessian matrix H in [I.22] and its principal minors $|H_{m+1}|, \dots, |H_{m+(n-m)}| \equiv |H|$. We can then state the condition: If \hat{x} satisfies the first order conditions and the bordered principal minors $|H_{m+j}| > 0$, $j = 1, \dots, n - m$, then \hat{x} yields a minimum of $f(x)$ subject to the constraints $g^k(x) = 0$, $k = 1, \dots, m$.

Again, it can be shown that these local conditions are concerned with the relative curvatures of the contours of the f and g^k functions in the neighbourhood of the optimal point. A *global* sufficient condition is that $f(x)$ be strictly quasi-concave and the $g^k(x)$ quasi-concave on their domains (or quasi-convex and strictly quasi-concave respectively).

Exercise 2I

1. In the example analysed in conditions [I.5]–[I.7], show that $\partial x_1 / \partial \alpha_1 = (-\lambda^* x_1^2 / D) - x_1^* (\partial x_1 / \partial \alpha_3)$. What is the sign of D ? If we cannot put signs to u_{22} and u_{12} , explain why we cannot sign the derivatives in [I.7]. If u is interpreted as a utility function, α_1 , α_2 as prices, and α_3 as income, relate the analysis here to the Slutsky equation in section 4B.
2. Develop second-order conditions for a minimum of a two-variable one-constraint problem along lines analogous to Fig. 2.20 and equation [I.18] of the text.

J. The Envelope Theorem

What we have called the 'standard' method of comparative statics is useful because it is routine, even programmable: one just follows the steps. However, in a problem of any

size it can be a very tedious method. Moreover, one may end up with a complicated expression in determinants which is not unambiguously signed. Considerable art and ingenuity may then be necessary to find and interpret economically interesting conditions under which a particular sign can be attached to a comparative statics effect. The more modern approach of *duality theory* applies the art and ingenuity at the outset to provide a more elegant and insightful mode of comparative statics analysis. Whenever possible, we adopt this approach in this book (see, for example, Chapters 4 and 9). A cornerstone of this approach is the *envelope theorem*: in fact, much of what we do will consist of repeated applications of this theorem in various contexts. Here we make clear its general form.

Suppose we have a classical maximization problem of the form

$$\max_x f(x, \alpha) \text{ s.t. } g^j(x, \alpha) = 0 \quad j = 1, \dots, m$$

where x is an $n(>m)$ -component vector and α is an l -vector of parameters. Let $v^* = f(x^*, \alpha)$ denote the value of the objective function at the optimal point. The Lagrange function is $L = f(x, \alpha) - \sum_{j=1}^m \lambda_j g^j(x, \alpha)$. Then the envelope theorem states:

$$\frac{\partial v^*}{\partial \alpha_k} = \frac{\partial L}{\partial \alpha_k} = f_{\alpha_k}(x^*, \alpha) - \sum_{j=1}^m \lambda_j^* g_{\alpha_k}^j(x^*, \alpha) \quad k = 1, \dots, l \quad [\text{J.1}]$$

That is, the effect of varying α_k on the optimised objective function is given by the partial derivative of the Lagrange function with respect to α_k , evaluated at the optimal solution point (x^*, λ^*) .

We have already seen one important application of this theorem, when we established the interpretation of the Lagrange multipliers in section 2G (refers to Chapter 2, section G. This abbreviation is used throughout.) We will see many more applications in the rest of this book and so we will not consider further examples here. We simply present the proof:

Differentiating v^* totally we have

$$dv^* = \sum_{i=1}^n f_i dx_i^* + f_{\alpha_k} d\alpha_k$$

Differentiating each constraint totally gives

$$dg^j = \sum_{i=1}^n g_i^j dx_i^* + g_{\alpha_k}^j d\alpha_k = 0 \quad j = 1, \dots, m$$

since $g^j = 0$ must continue to hold for any $d\alpha_k$. Multiplying through by λ_j^* , summing over j and noting that the result still equals zero allows us to write

$$\begin{aligned} dv^* &= \sum_{i=1}^n f_i dx_i^* + f_{\alpha_k} d\alpha_k - \sum_j \lambda_j^* \left[\sum_{i=1}^n g_i^j dx_i^* + g_{\alpha_k}^j d\alpha_k \right] \\ &= \sum_{i=1}^n \left[f_i - \sum_j \lambda_j^* g_i^j \right] dx_i^* + \left[f_{\alpha_k} - \sum_j \lambda_j^* g_{\alpha_k}^j \right] d\alpha_k \\ &= \left[f_{\alpha_k} - \sum_j \lambda_j^* g_{\alpha_k}^j \right] d\alpha_k \end{aligned}$$

which gives the result. The last step follows because, at the optimal point, $f_i - \sum_j \lambda_j^* g_j^i = 0$, $i = 1, \dots, n$, from the first-order conditions.

As we just noted, the special case in which α_k enters only one constraint, as a constraint constant, so that $g_{\alpha_k}^k = -1$, and $f_{\alpha_k} = g_{\alpha_k}^j = 0$, $k \neq j$, establishes that $dv^*/d\alpha_k = \lambda_k^*$. If α_k enters only the objective function, then $dv^*/d\alpha_k = f_{\alpha_k}(x^*, \alpha)$. This is also the form of the envelope theorem for the *unconstrained* maximization problem, $\max f(x, \alpha)$ (confirm). In this case the intuition behind the envelope theorem is straightforward. Changes in the parameter α_k alter the value of the objective function directly if f depends on α_k and indirectly because a change in α_k leads to changes in the optimal values of the choice variables x^* . However, the marginal value of changes in the choice variables is zero at the optimum and thus the indirect effects of α_k on v^* via x^* are also zero. Hence only the direct effect of α_k on the objective function matters.

K. Conclusions

In this chapter we have set out the central concepts of optimization theory and have given the main theorems. The purpose of this is to clarify the underlying structure of the microeconomic models we consider in the rest of this book. A good grasp of optimization theory should greatly increase the reader's ability to understand and extend these models.

Notes

1. This is not restrictive. Suppose the problem is to minimize some function $h(x)$. Then, this is equivalent to the problem of maximizing $f(x) = -h(x)$, since a solution to the latter problem solves the former. Making use of this, all the statements made in the present chapter about maximization problems can be applied directly to problems of minimization.
2. These conditions are sufficient, but not necessary, because a maximum may occur at a point x^* at which $f''(x^*) = 0$, so that proper necessary and sufficient conditions have to be stated in terms of higher order derivatives. For discussion of this see R. G. D. Allen (1938) *Mathematical Analysis for Economists*, Macmillan, Ch. XIV, and section 2I.
3. The fact that \bar{f} lies on the chord vertically above \bar{x} stems from the essential property of straight lines. Thus, given two points (x_1, y_1) and (x_2, y_2) , the points on the straight line joining them are given by:

$$\begin{aligned}(\bar{x}, \bar{y}) &= k(x_1, y_1) + (1-k)(x_2, y_2) \\ &= (kx_1 + (1-k)x_2, ky_1 + (1-k)y_2) \quad \text{for } 0 \leq k \leq 1\end{aligned}$$

In the present discussion we simply have $y_1 = f(x_1)$ and $y_2 = f(x_2)$ for some function $f(x)$.

4. There is unfortunately the possibility of a confusion in terminology. The reader may well have met a description of this type of contour as 'convex to the origin' which of course it is, and yet the function is *quasi-concave*. The reason for this term can be seen by comparing [B.10] and [B.4].
5. A function $h(x)$ is said to be quasi-convex, if $f(x) = -h(x)$ is quasi-concave. The reader should draw a contour of a quasi-convex function, on the assumption that all its partial

derivatives are strictly positive. Show in two dimensions that if the function $g(x_1, x_2)$ is quasi-convex, then the set of points satisfying: $g(x_1, x_2) \leq b$ is a convex set.

References and further reading

The reader with a good background in mathematics will find a rigorous treatment of the material presented in this chapter in

M. F. Intriligator, *Mathematical Optimisation and Economic Theory*. Prentice-Hall, Englewood Cliffs, NJ, 1971, chs. 2-4.

A. Takayama, *Mathematical Economics*, Cambridge University Press, Cambridge, 2nd edn, 1985, ch. 1.

A less heavily mathematical approach, which also suggests wider applications of the concepts of this chapter, can be found in:

T. Koopmans, *Three Essays on the State of Economic Science*, McGraw-Hill, London, 1957, ch. 1.

An excellent account of the structure, properties and uses of optimization problems in economics is provided in

A. Dixit, *Optimization in Economic Theory*, Oxford University Press, 2nd edn, 1990.

CHAPTER 3

The theory of the consumer

The central assumption in the theory of the consumer is that of optimization: given the feasible set of consumption bundles open to the consumer, he chooses the one he prefers. The purpose of the theory is first to characterize the bundle of goods which will be chosen, and second, to predict how the optimal choice will change in response to changes in the feasible set.

In analysing the consumer's optimal choice, we proceed in three steps. We first construct a model of the consumer's preferences, which allows us to specify certain important properties of the consumer's ranking of consumption bundles in terms of 'better', 'worse' or 'as good as'. We then examine how the prices of commodities in conjunction with the consumer's income (or his initial endowment of commodities in a more general model) together determine his feasible set of consumption bundles. Finally, by applying the model of the consumer's preference ordering to the feasible set, we are able to determine the characteristics of the optimal choice.

A. The preference ordering

A consumption bundle will be denoted by a vector:

$$x = (x_1, x_2, \dots, x_n)$$

where x_i , $i = 1, 2, \dots, n$, is the amount of the i th good in the bundle. Each x_i is assumed to be non-negative – the consumer can consume only zero or a positive quantity of each good – and also is taken to be perfectly divisible – goods do not come in lumpy discrete amounts.

The meaning of the terms 'preference' and 'indifference' is taken as understood; we take it for granted that everyone knows what is meant by the statement, 'I prefer this to that', or, 'I am indifferent between this and that'. In the present case, we assume that the consumer can make statements such as, 'I prefer consumption bundle x' to x'' ', or, 'I am indifferent between x' and x'' '. To put it more formally, we introduce the symbol \succsim which is read 'is preferred or indifferent to', or 'is at least as good as', or 'is no worse than' and

we let $x' \succsim x''$ stand for the statement that the consumer regards x' as at least as good as x'' , and $x'' \succsim x'$ for the converse. This symbol is called the *preference-indifference relation*.

Recall the view we have of the way the consumer chooses: he will rank the consumption bundles in the feasible set in order of preference, and choose one which comes highest in the ranking. This preference ranking can be thought of as being arrived at by repeated application of the preference-indifference relation to successive pairs of consumption bundles. For the purpose of our theory, we want the preference ranking to have certain properties, which give it a particular, useful structure. We build these properties up by making a number of assumptions, first about the preference-indifference relation itself, and then about some aspects of the preference ranking to which it gives rise. We now go on to examine these assumptions.

As a preliminary, suppose the consumer told us that:

$$x' \succsim x'' \quad \text{and} \quad x'' \succsim x'$$

in words, ' x' is preferred or indifferent to x'' ', and ' x'' is preferred or indifferent to x' '. Since we would regard him as talking nonsense – violating the meaning of the word 'preferred' – if he told us that x' is preferred to x'' and x'' is preferred to x' , this must mean that x' is indifferent to x'' . We write ' x' is indifferent to x'' ', as $x' \sim x''$. Suppose, alternatively, the consumer told us that:

$$x' \succ x'' \quad \text{and} \quad \text{not} \quad x'' \succ x'$$

This must mean that x' is preferred to x'' and this is written $x' \succ x''$. Thus we have as implications of the meaning of the preference-indifference relation:

- (a) $x' \succ x''$ and $x'' \succ x'$ implies $x' \sim x''$
- (b) $x' \succ x''$ and $\text{not } x'' \succ x'$ implies $x' \succ x''$.

We can now proceed to the assumptions which give the desired properties to the consumer's preference ordering.

Assumption 1. Completeness. For any pair of bundles x' and x'' , either $x' \succsim x''$ or $x'' \succsim x'$ (or both).

This assumption says in effect that the consumer is able to express a preference or indifference between any pair of consumption bundles however alike or unlike they may be. This ensures that there are no 'holes' in the preference ordering, points or areas to which it does not apply. It also implies that given some bundle x' , every other bundle can be put into one of three sets:

1. the set of bundles preferred or indifferent to x' , which is called the 'better set' for x' ;
2. the set of bundles indifferent to x' , which is called the indifference set of x' ;
3. the set of bundles to which x' is preferred or indifferent, which is called the 'worse set' for x' .

These sets, and especially 2, play an important part in what follows.

Assumption 2. Transitivity. For any three bundles x' , x'' , x''' , if $x' \succ x''$ and $x'' \succ x'''$ then $x' \succ x'''$.

Intuitively, this is a consistency requirement on the consumer. Given the first two statements, if the third did not hold, so that $x''' \succ x'$, we would feel there was an inconsistency in his preferences. The assumption has an important implication for the 'indifference sets' just defined, in that it implies that no bundle can belong to more than one such set. For suppose that $x' \sim x''$, so that x'' belongs to the indifference set of x' ; and also that $x'' \sim x'''$, so x'' belongs to the indifference set of x''' . If $x' \sim x'''$, then there is no problem, since all three bundles are in the same indifference set. But suppose $x''' \succ x'$. Then x'' must be in two indifference sets, that of x' and that of x''' . But then we have:

$$x' \sim x'' \quad \text{and} \quad x'' \sim x''' \quad \text{but} \quad x''' \succ x'$$

which violates the assumption of transitivity. Thus given this assumption, no bundle can belong to more than one indifference set. A way of putting this is to say: *The transitivity assumption implies that indifference sets have no intersection.*

Assumption 3. Reflexivity. $x' \succsim x'$.

In words, any bundle is preferred or indifferent to itself. Since we can interchange the two sides of the relation, the assumption has the implication that a bundle is indifferent to itself, which seems trivially true. However, its implication is less trivial: it ensures that every bundle belongs to at least one indifference set, namely that containing itself, if nothing else.

These three properties of the preference-indifference relation allow us to conclude that every bundle (completeness) can be put into one indifference set (reflexivity) and no more than one indifference set (transitivity). Thus we can *partition* any given set of consumption bundles, by use of the relation, into non-intersecting indifference sets, which provide us with a useful way of representing a particular preference ordering. The indifference sets can be ranked in order of preference on the basis of the ranking of the bundles they contain. The following assumptions we make about the consumer's preferences are chiefly designed to give these sets a particular structure.

Assumption 4. Non-satiation. A consumption bundle x' will be preferred to x'' if x' contains more of at least one good and no less of any other, i.e. if $x' \succ x''$.

This assumption establishes a relationship between the quantities of goods in a bundle and its place in the preference ordering – the more of each good it contains the better. Moreover, this is held to be true however large the amounts of the goods in the bundle, hence the term 'non-satiation' – the consumer is assumed never to be satiated with goods. This assumption is much stronger than we need to make in two respects. It first implies that none of the goods is in fact a 'bad', a commodity such as garbage or aircraft noise which one would prefer to have less of. Second, it assumes that the consumer is never satiated in any good. We could generalize by allowing some goods to be bads, and by assuming non-satiation only in at least one good, without changing anything of significance in the results of the theory. For simplicity, however, we adopt the stronger assumption here.

The non-satiation assumption has two important consequences for the nature of indifference sets, which are best expressed geometrically. In Fig. 3.1, x_1 and x_2 are goods,

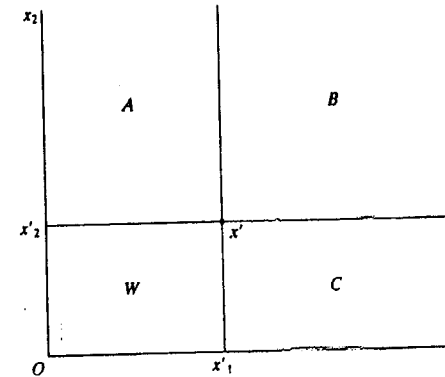


Fig. 3.1

and $x' = (x_1', x_2')$ is a consumption bundle. Because of assumption 4, all bundles in the area B (including the boundaries, except for x' itself) must be preferred to x' , and all points in the area W (again including the boundaries except for x') must be inferior to x' . The first consequence of the assumption is then that points in the indifference set for x' (if there are any besides x'), must lie in areas A and C. In other words, if we imagine moving between bundles in the indifference set, we can only do so by *substituting or trading off* the goods – giving more of one good must require taking away some of the other good in order to stay within the indifference set. The second consequence is that an indifference set is never 'wider' than a single point – its geometric representation can never be an area or band, though it may be a single point, an unconnected set of points or a curve. For suppose x' was contained in an indifference set which was a band. Then some bundles indifferent to it must lie in areas B and W, which violates assumption 4. Thus the assumption implies that an indifference set cannot be thick at point x' , or, by extending the argument, at any point contained in it.

None of the assumptions we have made so far, however, implies that there must be more than one point in an indifference set, or, if there is more than one point, that these make up a continuous line or curve. For example, as is shown in Appendix 1 to this chapter, the so-called *lexicographic ordering* satisfies assumptions 1 to 4, but its indifference sets each consist of only one point. We know that from the point of view of solving optimization problems continuity is a very important property (section 2C), and since we shall in effect be using indifference sets (or their geometric representation: indifference surfaces and, in the two-good case, indifference curves) to model the consumer's problem, it is a property we should like them to possess. Hence we make the assumption of continuity.

Assumption 5. Continuity. The graph of an indifference set is a continuous surface.

This implies that the surface, or curve in two dimensions, has no gaps or breaks at any point. In terms of the consumer's choice behaviour, what we are saying is this: given two goods in his consumption bundle, we can reduce the amount he has of one good, and however small this reduction is, we can always find an increase in the other good which will exactly compensate him, i.e. leave him with a consumption bundle indifferent to the first. The reader should confirm diagrammatically that this is possible only if the

indifference surface is everywhere continuous. (See Appendix 2 for a more formal treatment of this assumption.)

We now want to place some restrictions on the shape of the indifference surfaces or curves. From assumption 4 we already know that they must be negatively sloped, and so now we have to say something about their curvature. Recall the earlier definition of the better set of a point x' , as the set of bundles which are preferred or indifferent to x' . Then we make the assumption:

Assumption 6. Strict convexity. Given any consumption bundle x' , its better set is strictly convex.

Figure 3.2 illustrates for the two-good case. The better set for the point x' is the set of points on the indifference curve I' and in the shaded area, and this is drawn as strictly convex. There is an important technical reason for making this assumption: we know (from section 2E) that, given also that the feasible set is convex, the consumer's optimal point will as a result be a unique local – and therefore a global – optimum, and this is of considerable value when we come to analyse the consumer's responses to changes in the feasible set.

There is also a basis for the assumption in terms of economic behaviour, which can be expressed in two ways. From Fig. 3.2 it is clear that if we move the consumer along his indifference curve leftward from point x' , reducing the quantity of x_1 by small, equal amounts, we have to compensate, to keep him on the indifference curve, by giving him larger and larger increments of x_2 . In other words, the curvature implies that the smaller the amount of x_1 , and larger the amount of x_2 held by the consumer, the more valuable to him are marginal changes in x_1 relative to marginal changes in x_2 . It is argued that this is a common feature of consumer preferences.

A second way of rationalizing the curvature is as follows. In Fig. 3.2, $x' \sim x''$. Consider the straight line joining these two points. Any point on this line, for example \bar{x} , is a convex combination of x' and x'' , in that it can be expressed as:

$$\bar{x} = kx' + (1 - k)x'' \quad 1 \geq k \geq 0 \quad [\text{A.1}]$$

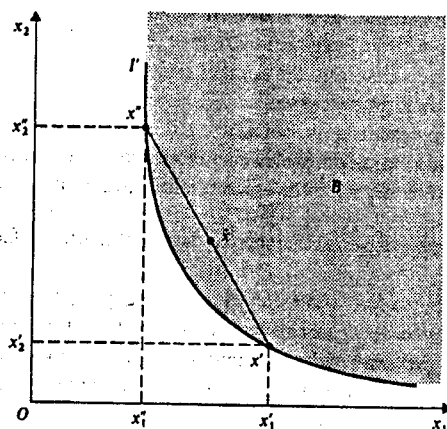


Fig. 3.2

i.e. the bundle \bar{x} contains an amount of x_1 given by $kx'_1 + (1 - k)x''_1$, and an amount of x_2 given by $kx'_2 + (1 - k)x''_2$. So, for example, if $k = \frac{1}{2}$, \bar{x} lies halfway along the line, and contains half of x'_1 plus half of x''_1 , and half of x'_2 plus half of x''_2 . We call such a convex combination a mixture of x' and x'' .

It follows from the strict convexity assumption that any mixture along the line will be preferred to x' and x'' (in fact this is the formal definition of strict convexity of the better set – see section 2B). Thus, the consumer always prefers a mixture of two consumption bundles which are indifferent to each other, to either one of those bundles. Again it is argued that this preference for mixtures is a commonly observed aspect of consumer behaviour.

A weaker convexity assumption than assumption 6 can be made: we could assume that the better set is convex but not strictly convex. This means that we allow the possibility of linear segments in the indifference curves, as Fig. 3.3 illustrates. The better sets for points x' , x'' and x''' respectively are each convex but none is strictly convex. Linearity in the indifference curve over some range implies that within this range, the valuation of marginal decreases in one good relative to marginal increments in the other remains constant – successive equal reductions in the amount of one good are compensated by successive equal increases in the amount of the other. Alternatively, a mixture of two indifferent bundles, in the sense just defined, is indifferent to the two, rather than preferred to them. The reason for excluding such linearity by the strict convexity assumption is, as we shall see, to ensure that the solution to the consumer's problem is a unique point and not a set of infinitely many points.

As a result of these six assumptions, we can represent the preference ordering of the consumer by a set of continuous convex-to-the-origin indifference curves or surfaces, such that each consumption bundle lies on one and only one of them. Moreover, as a result of assumption 4 we can say that bundles on a higher indifference surface are preferred to those on a lower. Thus, the best consumption bundle open to a consumer is the one lying on the highest possible indifference surface. We therefore have part of the analytical apparatus we need to solve the consumer's choice problem. It is, however, of interest to

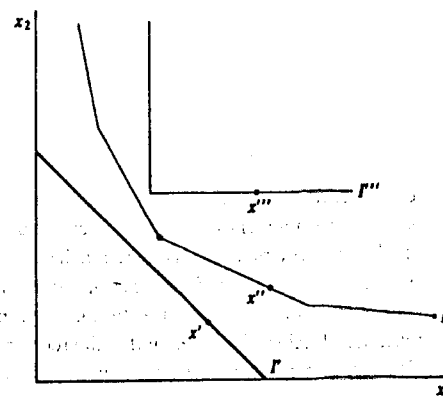


Fig. 3.3

examine at this point a somewhat different way of representing the consumer's preference ordering.

The utility function

Historically the word 'utility' was used in economics to denote the subjective sensations – satisfaction, pleasure, wish-fulfilment, cessation of need, etc. – which are derived from consumption, and the experience of which is the object of consumption. The economists in the late nineteenth century who were concerned with constructing a theory of consumer choice went further than this definition and regarded utility as something which could be measured as an absolute quantity, in the same way as, say, weight can be measured. They thought it possible to speak of the total quantity of utility derived from consuming a given bundle of goods, of subtracting such quantities from each other, and discussing how these differences changed as consumption varied. Thus was developed the 'law of diminishing marginal utility'. However, even then some of these economists were unhappy about this measurability and it came increasingly under attack as the theory developed. The position which is generally accepted now is that the subjective sensations grouped under the name 'utility' are not capable of being treated as quantities in this sense. An important reason for the adoption of that position was the demonstration that for the purpose of constructing a theory of consumer choice, not only the measurement of utility, but the very concept itself, is unnecessary. As we have seen, we can base a theory of choice on the concepts of preference and indifference, and *nothing more is needed* for the theory than the set of indifference curves (or surfaces) with their assumed properties.

However, for some methods of analysis it is useful to have a function which provides a numerical representation of the preference ordering. That is, it is useful to have a rule for associating with each consumption bundle a real number which indicates its place in the ranking. The reason for this is that we can then apply the standard method of constrained maximization of a function to obtain the solution to the consumer's choice problem.

A suitable rule of association or function can be defined in the following way. On the assumptions made about the consumer's preferences we can partition the consumption bundles into indifference sets and can rank these sets. A rule or function $u(x)$ which assigns a real number u to each bundle x is said to *represent* the consumer's preferences if all bundles in the same indifference set have the same number and bundles in preferred indifference sets have higher numbers, i.e.

(a) $u(x') = u(x'')$ if and only if $x' \sim x''$

(b) $u(x') > u(x'')$ if and only if $x' \succ x''$

Any function satisfying these simple requirements is a *utility function* for the consumer.

A utility function is merely a way of attaching numbers to the consumer's indifference sets such that the numbers increase as higher or more preferred sets are reached. It reflects the *ordering* of the bundles by the consumer and so is an *ordinal* function. Since we only require that the consumer can rank bundles and the utility function is a numerical representation of this ordering, no significance attaches to the size of the difference between numbers attached to different bundles. We are concerned only with the *sign* of the difference, i.e. whether $u(x') \geq u(x'')$ or whether x' is preferred or indifferent to x'' or x'' preferred to x' .

There are an infinite number of ways of attaching numbers to bundles which are consistent with the requirements (a) and (b) above: the utility function is not unique. For example, given four consumption bundles x', x'', x''', x'''' , any one of the columns in the following table is an acceptable numerical representation of the preference ordering:

	$u(x)$	$v(x)$	$w(x)$
x'	3	10000	500
x''	3	10000	500
x'''	2	2	499
x''''	1	1.5	1.9

where $v(x)$ and $w(x)$ denote functions which obey the rule in (a)–(b) above, but which differ from $u(x)$. To put this more formally, we could regard the function $v(x)$ as being derived from $u(x)$ by applying, at each x , some *rule of transformation*, such as, for example, 'When $x = x'$ multiply $u(x')$ by 3333 to obtain $v(x)$ '. That is in general we write:

$$v(x) = T[u(x)] \quad [A.2]$$

where $T[]$ denotes the rule of transformation we devise. The only restriction we place on this transformation rule is that when u increases, v must increase, because then v will correctly represent the preference ordering. Such a transformation is called '*positive monotonic*', because v must always increase with u . Hence, we say that the function $u(x)$ is *unique up to a positive monotonic transformation* meaning that we can always derive another permissible representation of the preference ordering by applying some positive monotonic transformation T to $u(x)$. Examples of such transformations are:

$$v(x) = [u(x)]^2 \quad [A.3]$$

$$v(x) = 3 + 2u(x)$$

$$v(x) = 5 + \log u(x)$$

where the transformation T is defined by a simple function. As the table above showed, we do not *have* to define T in such a simple way.

So far we have taken it for granted that a function $u(x)$ which gives a numerical representation of a preference ordering actually does exist. This is something we should consider explicitly. What do we have to assume in order to ensure that the function exists? Consider first assumptions 1–3 above, on completeness, transitivity and reflexivity. Recall that they resulted in a family of indifference sets such that every consumption bundle belonged to one and only one set. We might then reason intuitively that since the $u(x)$ function effectively assigns numbers to indifference sets, there can be no problem. We would, however, be wrong. It can be shown that we may have a preference ordering satisfying assumptions 1–3 (and 4), but for which no numerical representation exists – we cannot apply to it the rule for assigning numbers to consumption bundles that we set out earlier. An ordering for which this is true is the lexicographic ordering discussed in Appendix 1. The existence of this counter-example tells us that assumptions 1–4 are not sufficient to guarantee existence of a numerical representation of a preference ordering. The further assumption which solves the problem is that of continuity. It can be shown (see Appendix 2) that if assumption 5 holds, so that the indifference surfaces are continuous, a continuous numerical representation $u(x)$ can always be constructed for the preference ordering.

We can now consider the relation between the function $u(x)$ and the indifference sets, which are the fundamental expressions of the consumer's preference ordering. Consider the set of consumption bundles which satisfy:

$$u(x) = u^0 \quad [\text{A.4}]$$

where u^0 is some given number. Since these consumption bundles yield the same value of the function they must constitute an indifference set. A set of values of the independent variables in a function which yield a constant value of the function is said to define a *contour* of that function. Hence the indifference sets are contours of the function $u(x)$, and the assumptions 4 and 6 which define the shape of the indifference sets can just as well be interpreted as defining the properties of the contours of $u(x)$. This implies that $u(x)$ is what we called in section 2B, a *strictly quasi-concave* function. In addition, we know that a consumption bundle which yields a higher value of the function than another will always be preferred, and so we can interpret the desire to choose the preferred alternative in some given set of alternatives as equivalent to maximizing the function $u(x)$ over that set. Thus we can represent the consumer's choice problem as one of constrained maximization of a strictly quasi-concave function.

In formulating the consumer's choice problem in this way, it is useful if we can use methods of differentiation to find solutions. The assumptions made so far do not imply differentiability: for example, Fig. 3.4 shows a contour which satisfies all the assumptions but is not differentiable at x' – the slope of the contour is not uniquely defined at that point, which is a corner. To rule out such cases, we make the assumption of differentiability (since differentiability implies continuity, we could regard assumption 7 as replacing assumption 5).

Assumption 7. Differentiability. Utility functions are differentiable to any required order.

This assumption rules out cases in which the slope of an indifference surface or curve makes a sudden jump, as in Fig. 3.4. We now have to examine more closely the interpretation of the slope of an indifference curve.

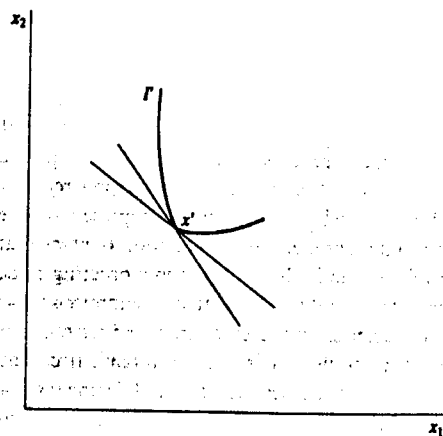


Fig. 3.4

Recall that in discussing assumption 6 we used the idea of successive small reductions in x_1 being compensated by small increments in x_2 just enough to stay on the indifference curve. This can be thought of as defining a 'required rate of compensation', whose (absolute) value increases as we move leftward along the indifference curve. As usual with ratios of *finite* changes, there is an ambiguity arising out of the arbitrariness of the size of the change, and so we find it useful to go to the limit and define the derivative:

$$\left. \frac{dx_2}{dx_1} \right|_{u \text{ constant}} = \lim_{\Delta x_1 \rightarrow 0} \left(\frac{\Delta x_2}{\Delta x_1} \right) \quad [\text{A.5}]$$

where the notation on the left-hand side is intended to emphasize that we are constraining the changes in x_1 and x_2 to keep a constant value of the function u . In effect, we view the indifference curve as defining x_2 as a function of x_1 , which could be called an 'indifference function' or 'contour function'. Then the derivative we have defined above is the slope of this function at a point. Figure 3.5 illustrates. The slope of the tangent L to the indifference curve at x' gives the value of the above derivative at x' . As we take points leftward along the indifference curve, the absolute value of the derivative increases. The figure also shows a sequence of finite changes; the ratio $\Delta x_2/\Delta x_1$ gives the *average* rate of change of x_2 with respect to x_1 over an arc of the curve, and its value will depend on the size of the change Δx_1 .

Important derivatives in economics are always called the *marginal* something or other, and this is no exception. We define the *marginal rate of substitution* of good 2 for good 1, written MRS_{21} , as:

$$MRS_{21} = - \left. \frac{dx_2}{dx_1} \right|_{u \text{ constant}} \quad [\text{A.6}]$$

The negative sign occurs because we wish MRS_{21} to be positive. Assumption 6 implies that MRS_{21} varies inversely with x_1 . We also define the *marginal rate of substitution* of

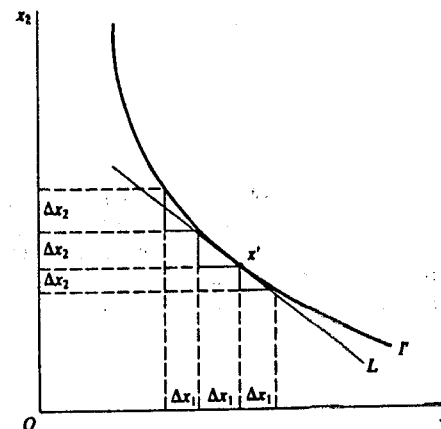


Fig. 3.5

good 1 for good 2, written MRS_{12} , as:

$$MRS_{12} = - \left. \frac{dx_1}{dx_2} \right|_{u \text{ constant}} \quad [A.7]$$

which refers to the slope of an indifference curve relative to the x_2 -axis. Since the two are reciprocals of each other, it is enough to work always with just one of them, and MRS_{21} will usually be taken.

Along an indifference surface we have:

$$u(x) = u^0$$

where u^0 is a given constant. Assumption 7 allows us to differentiate u totally to obtain:

$$du = u_1 dx_1 + u_2 dx_2 + \dots + u_n dx_n = 0 \quad [A.8]$$

where u_i , $i = 1, 2, \dots, n$, is the partial derivative $\partial u / \partial x_i$ or the *marginal utility* of good i . This equation constrains the differentials dx_i to be such as to maintain $u = u^0$. Let us assume that the quantities of all goods other than the first two are held constant, i.e. $dx_i = 0$, $i = 3, \dots, n$. Then, by rearranging we have:

$$\left. - \frac{dx_2}{dx_1} \right|_{u \text{ constant}} = \frac{u_1}{u_2} = MRS_{21} \quad [A.9]$$

Thus the marginal rate of substitution at a point can be expressed as the ratio of marginal utilities at that point. Since u_1 and u_2 are in general functions of all n goods, so is MRS_{21} . Clearly, the above procedure can be used to derive the marginal rate of substitution for any pair of goods.

Note that useful though it is to have this relationship between marginal rates of substitution and partial derivatives of $u(x)$, in a sense it is the former which are more fundamental. The preference ordering of the consumer uniquely determines the indifference sets and hence the marginal rates of substitution. The partial derivatives on the other hand, depend on the particular function used to represent the consumer's preferences, i.e. to label the indifference sets.

Properties of marginal utility

If x_i increases with the amounts of all other goods held constant the consumer achieves a better bundle and hence the utility number must increase, so that marginal utility of the i th good is positive: $u_i(x) > 0$. The *sign* of the marginal utility of a good is the same for all numerical representations of the consumer's preferences (i.e. for all utility functions) but the size of the marginal utility is not. If u is a utility function and $v = T[u(x)]$ is a transformation of u with the property that $T' = dT/du > 0$ then $v(x)$ is also a utility function. The partial of v with respect to x_i is

$$\frac{\partial v}{\partial x_i} = v_i = T' \cdot \frac{\partial u}{\partial x_i} = T' u_i \quad [A.10]$$

and, since by assumption $T' > 0$, the sign of v_i is the sign of u_i but $v_i \neq u_i$.

The rate of change of marginal utility of x_i with respect to x_i is the second partial derivative of u with respect to x_i : $u_{ii} = \partial^2 u / \partial x_i^2$. Neither the sign nor the magnitude of the rate of change of u_i are the same for all representations of preferences. For example, with the function v considered in the previous paragraph.

$$v_{ii} = \frac{\partial^2 v}{\partial x_i^2} = \frac{\partial}{\partial x_i} (T' u_i) = T'' u_{ii} + T' u_{ii}$$

Hence the sign of v_{ii} is the same as the sign of u_{ii} for all T only if $T'' = d^2 T / du^2 = 0$, but the only restriction on T is that $T' > 0$. Statements about increasing or diminishing marginal utility are therefore meaningless, because we can always find a function to represent the consumer's preferences which contradicts the statement.

[A.9] makes the important point that ratios of marginal utilities are invariant to permissible transformations of the utility function since they must all equal the marginal rate of substitution, which is determined by the consumer's preferences. Using the utility functions u and v above and [A.10], we see that

$$MRS_{ij} = \frac{v_j}{v_i} = \frac{T' \cdot u_j}{T' \cdot u_i} = \frac{u_j}{u_i} \quad [A.11]$$

Our warnings about the meaninglessness of statements about the size of changes in utility are valid for the preferences which satisfy the assumptions of this chapter but, as we will see in Chapter 19, if certain additional restricting assumptions about an individual's preferences are made, it becomes sensible to talk of the rate of change of marginal utility. These extra assumptions are unnecessary for our present purposes and so we do not adopt them until they are needed, when we study decision-making under conditions of uncertainty.

Exercise 3A

1. Show that if indifference curves intersect the consumer is inconsistent.
2. Construct a set of indifference curves which satisfy all the assumptions of this section, *except*:
 - (a) one of the 'goods' is in fact a bad; or
 - (b) the consumer may reach a point at which he is satiated with one good but not the other; or
 - (c) the consumer may reach a point at which he is satiated with both goods (a 'bliss point'); or
 - (d) there is a quantity for each good up to which it is a good, and beyond which it is a bad.

Give concrete examples of goods which may fit each case.
3. Discuss the relationship between the non-satiation assumption and the idea of scarcity which underlies microeconomics.
4. Draw indifference curves relating to:
 - (a) red and blue matches with identical incendiary properties;

(b) left and right shoes of the same size, quality, design, etc.;

and state whether the corresponding utility function is strictly quasi-concave. Comment on the way in which the MRS_{21} varies along these indifference curves.

5. Mr A's indifference curves for water and diamonds satisfy the assumption of strict convexity, and he is endowed with a great deal of water and very few diamonds. Which of the following does this imply?

- (a) diamonds are more valuable to him than water;
- (b) he would give up a lot of water to get one more diamond;
- (c) he would give up more water for an extra diamond than would be the case if he had a combination, indifferent to the first, of less water and more diamonds.

B. The feasible set

We initially assume that the consumer has a given money income M , that he faces constant prices for all of the goods in his utility function and that he cannot consume negative quantities of any good. Then, recalling section 2A, the consumer's feasible set defined by these assumptions is the set of bundles satisfying

$$p_1x_1 + p_2x_2 + \dots + p_nx_n = \sum p_ix_i \leq M \quad [B.1]$$

$$x_1 \geq 0, x_2 \geq 0, \dots, x_n \geq 0$$

where p_i is the price of good i .

The feasible set in the two-good case is shown in Fig. 3.6 as the triangular area $Ox_1^0x_2^0$. $x_1^0 = M/p_1$ is the maximum amount of x_1 that can be bought with income M at a price of p_1 . x_2^0 is analogously defined. The budget constraint is $p_1x_1 + p_2x_2 \leq M$ in this two-good case, or:

$$x_2 \leq (M - p_1x_1)/p_2 \quad [B.2]$$

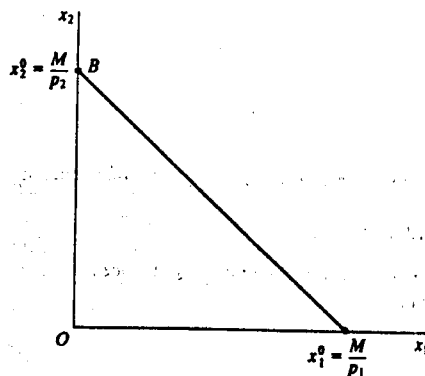


Fig. 3.6

which is satisfied by all points on or below the line B from x_1^0 to x_2^0 . B , the upper boundary of the feasible set, is known as the consumer's *budget line* and is defined by

$$x_2 = (M - p_1x_1)/p_2 \quad [B.3]$$

The slope of the budget line is therefore

$$\left. \frac{dx_2}{dx_1} \right|_{M \text{ constant}} = -\frac{p_1}{p_2} \quad [B.4]$$

where the notation on the left-hand side is to remind us that this is the rate at which a consumer with fixed income can exchange x_1 for x_2 on the market. A one-unit reduction in purchases of x_1 reduces expenditure by p_1 , and so, since 1 unit of x_2 costs p_2 , the consumer can buy p_1/p_2 extra units of x_2 . Therefore 1 unit of x_1 exchanges for p_1/p_2 units of x_2 .

As a preparation for the next section let us examine the consumer's feasible set in terms of the concepts introduced in section 2B. The feasible set is:

- (a) *bounded*, from below by the non-negativity constraints on the x_i and from above by the budget constraint, provided that M is finite and no price is zero. If, for example, $p_1 = 0$ then the budget line would be a line parallel to the x_1 axis through the point $x_2^0 = M/p_2$, and the feasible set would be unbounded to the right: since x_1 would be a free good the consumer could consume as much of it as he wished.
- (b) *closed*, since any bundle on the budget line B or the quantity axes is available.
- (c) *convex*, since for any two bundles x' and x'' in the feasible set, any bundle \bar{x} lying on a straight line between them will also be in the feasible set. Since \bar{x} lies between x' and x'' , and they both satisfy the non-negativity constraints, \bar{x} will also satisfy these constraints. \bar{x} will cost no more than the consumer's income: lying between x' and x'' it must cost no more than the more expensive of them, say x' . But since x' lies within the feasible set, so must \bar{x} . Hence \bar{x} is in the feasible set.
- (d) *non-empty*: provided that $M > 0$ and at least one price is finite the consumer can buy a positive amount of at least one good.

We will consider here the effects of changes in M and p_i on the feasible set, in preparation for section D where we examine their effects on the consumer's optimal choice. If money income increases from M_0 to M_1 , the consumer's feasible set expands as the budget line moves outward parallel with its initial position, as in Fig. 3.7(a). With $M = M_0$ the intercepts of the budget line B_0 on the x_1 and x_2 axes respectively are M_0/p_1 and M_0/p_2 and with $M = M_1$ they are M_1/p_1 and M_1/p_2 . A doubling of M for example, will double the value of the intercepts, since $M_1/p_2 = 2M_0/p_2$ when $M_1 = 2M_0$. The slope of the budget line is $-p_1/p_2$ and this is unaffected by changes in M .

Consider next an increase in p_1 , as shown in Fig. 3.7(b). Since M and p_2 are unchanged the budget line will still have the same M/p_2 intercept on the x_2 axis. An increase in p_1 will cause the budget line to pivot about M/p_2 and become more steeply sloped as p_1/p_2 becomes larger. In Fig. 3.7(b) a rise in p_1 to p'_1 shifts the x_1 intercept from M_0/p_1 to M_0/p'_1 where $M_0/p_1 > M_0/p'_1$ since $p_1 < p'_1$.

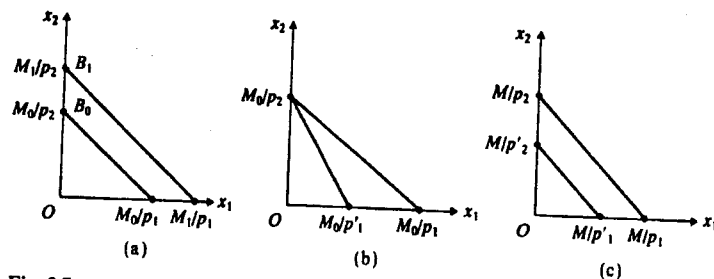


Fig. 3.7

Equal proportionate increases in all prices will cause the budget line to shift inwards towards the origin as in Fig. 3.7(c). Suppose p_1 and p_2 increase from p_1 and p_2 to kp_1 and kp_2 where $k > 1$. Then the slope of the new budget line is unchanged: $-kp_1/kp_2 = -p_1/p_2$, and the new intercepts are $M/kp_1 < M/p_1$ and $M/kp_2 < M/p_2$.

Finally, if all prices and M change in the same proportion the budget line is unchanged. The intercept on the i th axis after all prices and M change by the factor k is $kM/kp_i = M/p_i$ so the intercept is unaffected, as is the slope, which is $-kp_1/kp_2 = -p_1/p_2$.

Exercise 3B

- 1.* Suppose that the price of one of the commodities bought by the consumer rises as he buys larger quantities. What effect will this have on his feasible set? What interpretation can be given to the slope of his budget line? Can you show in the diagram the relationship between the average price (expenditure divided by quantity bought) and the marginal price?
- 2.* Many public utilities sell their products on multi-part tariffs. The consumer must pay a connection charge for the right to consume (say) electricity, irrespective of the amount consumed. The price paid for the first n units will exceed the price paid for any units consumed in excess of n . Draw the feasible set for the consumer, with electricity on one axis and a consumption good on the other. Distinguish between the average and marginal prices of electricity and investigate the effects of changes in the connection charge and the price of electricity.
3. Draw the feasible set of the consumer in Question 2a, Exercise 3A, assuming that the 'bad' is garbage and that there is a given price per bag of garbage removed, and a given amount of garbage produced per period by the consumer.

C. The consumption decision

Given the assumptions of the previous two sections, the consumer's problem of choosing the most preferred bundle from those available to him can be formally stated as

$$\begin{aligned} \max_{x_1, \dots, x_n} & u(x_1, x_2, \dots, x_n) \\ \text{s.t. } & \sum_i p_i x_i \leq M; \quad x_i \geq 0 \quad (i = 1, \dots, n) \end{aligned} \quad [\text{C.1}]$$

We can derive the equilibrium conditions which the solution to this problem must satisfy by a diagrammatic analysis of the two-good case. We leave to the latter part of this section a brief confirmation of our results using the more rigorous methods of Chapter 2.

From the assumptions of section A we can represent the consumer's preferences by a utility function which has indifference curves or contours like those of Fig. 3.8. All commodities are assumed to have positive marginal utility so that bundles on higher indifference curves are preferred to those on lower indifference curves. This assumption (a consequence of assumption 4 in section A) also means that the consumer will spend all his income since he cannot be maximizing u if he can buy more of some good with positive marginal utility. The consumer will therefore choose a bundle on his budget line B .

In Fig. 3.8 there is a *tangency solution* where the optimal bundle x^* is such that the highest attainable indifference curve I_1 is tangent to the budget line and the consumer consumes some of both goods. The slope of the indifference curve is equal to the slope of the budget line at the optimum:

$$\left. \frac{dx_2}{dx_1} \right|_{u \text{ constant}} = \left. \frac{dx_2}{dx_1} \right|_{M \text{ constant}}$$

The negative of the slope of the indifference curve is the marginal rate of substitution MRS_{21} ; and the negative of the slope of the budget line is the ratio of the prices of x_1 and x_2 . Hence the consumer's equilibrium condition can be written as

$$MRS_{21} = \frac{u_1}{u_2} = \frac{p_1}{p_2} \quad [\text{C.2}]$$

The consumer is in equilibrium (choosing an optimal bundle) when the rate at which he can substitute one good for another on the market is equal to the rate at which he is just content to substitute one good for another.

We can interpret this property of the optimal choice in a somewhat different way. If the consumer spent an extra unit of money on x_1 , he would be able to buy $1/p_1$ units of x_1 . $u_1 \Delta x_1$ is the gain in utility from an additional Δx_1 units of x_1 . Hence u_1/p_1 is the gain in

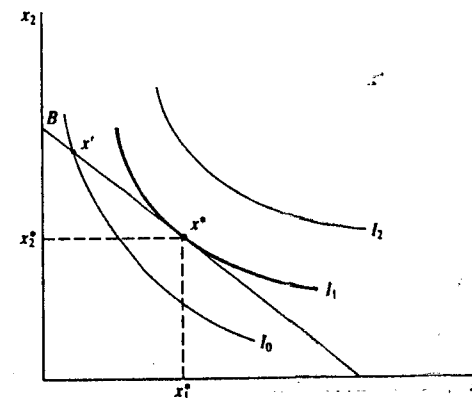


Fig. 3.8

utility from spending an additional unit of money on x_1 . u_2/p_2 has an analogous interpretation. The consumer will therefore be maximizing his utility when he allocates his income between x_1 and x_2 so that the marginal utility of expenditure on x_1 is equal to the marginal utility of expenditure on x_2 :

$$\frac{u_1}{p_1} = \frac{u_2}{p_2} \quad [\text{C.3}]$$

This is exactly the condition obtained by multiplying both sides of [C.2] by u_2/p_1 .

If the consumer's income were increased by a small amount he would be indifferent between spending it on x_1 or x_2 : in either case his utility would rise by $u_1/p_1 = u_2/p_2$. Hence, if we call the rate at which the consumer's utility increases as his income increases the marginal utility of income, denoted by u_M , we have

$$\frac{u_1}{p_1} = \frac{u_2}{p_2} = u_M \quad [\text{C.4}]$$

A more plausible optimum when there are many goods would be a *corner point solution*, where the optimal bundle x^* does not contain positive amounts of all goods, as in Fig. 3.9 where no x_2 is purchased. In this case the indifference curve at x^* is steeper than the budget line, i.e. has a smaller slope (remembering that the indifference curve and the budget line are negatively sloped).

Hence

$$\left. \frac{dx_2}{dx_1} \right|_{u \text{ constant}} < \left. \frac{dx_2}{dx_1} \right|_{M \text{ constant}} \quad [\text{C.5}]$$

and therefore

$$\left. \frac{-dx_2}{dx_1} \right|_{u \text{ constant}} = \text{MRS}_{12} = \frac{u_1}{u_2} > \frac{p_1}{p_2} = \left. \frac{-dx_2}{dx_1} \right|_{M \text{ constant}} \quad [\text{C.6}]$$

Rearranging, this equilibrium condition can be written

$$u_M = \frac{u_1}{p_1} > \frac{u_2}{p_2} \quad [\text{C.7}]$$

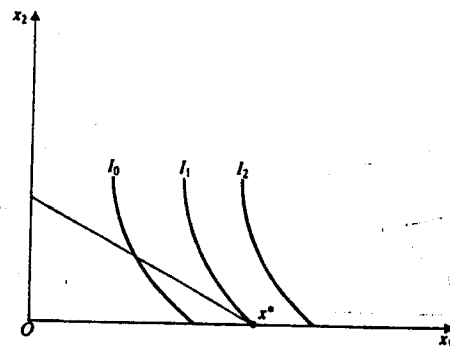


Fig. 3.9

The marginal utility of expenditure on the good purchased, x_1 , is greater than the marginal utility of expenditure on x_2 , the good not purchased. Because of the higher marginal utility of expenditure on x_1 than on x_2 the consumer would like to move further down the budget line substituting x_1 for x_2 but he is restrained by the fact that he cannot consume negative amounts of x_2 .

A more formal analysis

Since the consumer's preferences satisfy the assumptions of section A, the objective function in problem [C.1] above is continuous and strictly quasi-concave. From section B we see that the feasible set for the problem, defined by the budget and non-negativity constraints will be non-empty, closed, bounded and convex. From the Existence, Local-Global and Uniqueness Theorems of Chapter 2 the consumer's optimization problem will have a unique solution and there will be no non-global local solutions.

Since there is at least one good with positive marginal utility the consumer will spend his entire income and hence the budget constraint can be written as an equality constraint: $M - \sum p_i x_i = 0$. If we assume that the solution will be such that some of all goods will be consumed ($x_i^* > 0$ ($i = 1, \dots, n$) where x_i^* is the optimal level of x_i), then the non-negativity constraints are non-binding and we have a problem to which can be applied the method of Lagrange outlined in section 2G. The Lagrange function derived from [C.1] is

$$L = u(x_1, \dots, x_n) + \lambda [M - \sum p_i x_i] \quad [\text{C.8}]$$

and the first-order conditions for a solution to [C.1] are

$$\frac{\partial L}{\partial x_i} = u_i - \lambda p_i = 0 \quad (i = 1, \dots, n) \quad [\text{C.9}]$$

$$\frac{\partial L}{\partial \lambda} = M - \sum p_i x_i^* = 0 \quad [\text{C.10}]$$

If [C.9] is rewritten as $u_i = \lambda p_i$ and the condition on good i is divided by that on good j , we have

$$\frac{u_i}{u_j} = \frac{p_i}{p_j} \quad [\text{C.11}]$$

or: the marginal rate of substitution between two goods is equal to the ratio of their prices as in condition [C.3] above. Alternatively, [C.9] can be rearranged to give

$$\frac{u_1}{p_1} = \frac{u_2}{p_2} = \dots = \frac{u_n}{p_n} = \lambda \quad [\text{C.12}]$$

which is of course the n -good extension of the condition [C.4] derived earlier.

In section 2G we demonstrated that the value of the Lagrange multiplier λ was the rate at which the objective function increased as the constraint parameter was increased. In this case the objective function is the utility function and the constraint parameter is the

individual's money income so that λ is the rate at which utility increases as money income increases:

$$\lambda = \frac{du^*}{dM} = u_M \quad [\text{C.13}]$$

The Lagrange multiplier can be interpreted as the marginal utility of money income. This interpretation is supported by [C.12] since, as we argued above, u_i/p_i is the rate at which utility increases as more money is spent on good i .

Corner solutions

If the assumption that $x_i^* > 0$ for all i is dropped, the first order conditions for [C.1] are derived by maximization of the Lagrangean [C.8] subject to the direct non-negativity constraints on the choice variables. Reference to section 2H indicates that the conditions which must be satisfied by a solution to C.1 are

$$\frac{\partial L}{\partial x_i} = u_i - \lambda^* p_i \leq 0, \quad x_i^* \geq 0, \quad x_i^* \cdot (u_i - \lambda^* p_i) = 0$$

$$i = 1, 2, \dots, n \quad [\text{C.14}]$$

plus condition [C.10].

If [C.14] is rearranged to give

$$\lambda^* \geq \frac{u_i}{p_i}, \quad x_i^* \geq 0, \quad x_i^* \cdot \left(\frac{u_i}{p_i} - \lambda^* \right) = 0 \quad [\text{C.15}]$$

it can be given a straightforward economic interpretation: if the marginal utility of expenditure on good i , (u_i/p_i) , is less than the marginal utility of money at the optimal point, λ^* , then good i will not be bought since the consumer will get greater utility by expenditure on other goods. The same result can be derived from [C.7], where $x_i = 0$, since $u_i/p_i = u_M > u_2/p_2$, or $u_2 - u_M p_2 < 0$.

Exercise 3C

- 1.* If a consumer buys electricity on a multi-part tariff, as in Question 2, Exercise 3B, are the conditions of the Existence, Local-Global and Uniqueness Theorems satisfied?
- 2.* Derive and interpret the equilibrium conditions for the types of preferences postulated in Questions 2 and 4, Exercise 3A.
(Note: what must be assumed about the price of a 'bad'?)
3. Explain why a consumer would:
 - (a) not choose a point inside his budget line;
 - (b) not choose the bundle x^* in Fig. 3.8.

- 4.* Suppose that, as well as paying a price per unit of a good, the consumer has to pay a 'transactions cost' for using a market. Analyse the implications for the consumer's optimal choice of assuming:
 - (a) the transactions cost is paid as a lump sum;
 - (b) the transactions cost is proportional to price but independent of the quantity bought;
 - (c) the transactions cost is charged per unit of the good bought, but decreases the greater the amount bought.

5. Is the marginal utility of money income, λ^* or u_M , uniquely defined?

D. The comparative statics of consumer behaviour

The solution to the consumer's optimization problem depends on his preferences, the prices he faces and his money income. We can write this solution, which we call his *demand for goods*, as a function of prices and money income.

$$x_i^* = D_i(p_1, p_2, \dots, p_n, M) = D_i(p, M) \quad (i = 1, \dots, n) \quad [\text{D.1}]$$

where $p = (p_1, p_2, \dots, p_n)$ is the vector of prices, and the form of the *Marshallian demand function* D_i depends on the consumer's preferences.

The discussion in Chapter 2 on the properties of feasible sets and the objective function, and the assumptions made about those defined in this chapter, enable us to place restrictions on the form of the demand functions. First, provided that p, M are finite and positive, the optimization problem must have a solution, since the requirements of the Existence Theorem are satisfied. Second, the differentiability of the indifference curves and the linearity of the budget constraint imply that the optimal bundle will vary continuously in response to changes in prices and income, and that the demand functions are *differentiable*. Third, the conditions of the Uniqueness Theorem are satisfied and so the demand relationships are functions rather than correspondences. A *unique* bundle is chosen at each (p, M) combination.

Prediction of the bundle chosen in a given situation will require precise knowledge of the consumer's preferences, and hence our results so far are not very useful if we wish to test the model. We now consider the *comparative statics* properties of the model to see if they yield predictions which do not require information we do not have. We wish to investigate the effects of changes in the exogenous variables (prices, money income) on the equilibrium values of the endogenous variables (the consumer's demand for goods). In other words we want to predict what happens to the optimal bundle $x^* = (x_1^*, x_2^*, \dots, x_n^*) = (D_1, D_2, \dots, D_n)$ as the feasible set varies.

We consider first changes in the consumer's money income. In Fig. 3.10 B_1 is the initial budget line, x^* the initial bundle chosen. An increase in M , with p_1, p_2 constant, will shift the budget line outward parallel with itself, say to B_2 where x' is chosen. A further increase in M will shift the budget line to B_3 where x'' is chosen. The *income consumption curve* is the set of optimal points traced out as income varies in this way, with prices constant. In the case illustrated both x_1 and x_2 are *normal goods*, for which demand increases as money

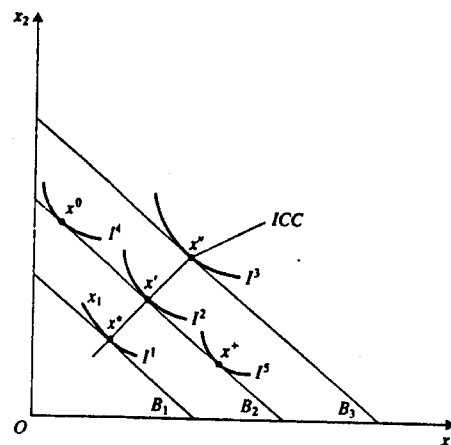


Fig. 3.10

income rises. However, with different preferences the consumer might have chosen x^0 or x^+ on B_2 . If x^0 had been chosen (if I^4 and not I^2 had been the consumer's indifference curve) then the demand for x_1 would have fallen as money income rose. x_1 would then be known as an *inferior good*. It is clear that a rise in M may lead to a rise, a fall, or no change in the demand for a good. Without knowledge of preferences we cannot predict whether a particular good will be inferior or normal. The theory of consumer behaviour cannot be tested by considering the effect of changes in M on the demand for a single good, since any effect is compatible with the theory.

The theory does predict, however, that *all* goods cannot be inferior. If the consumer reduces his demand for all goods when his income rises he will be behaving inconsistently. To show this, let x^* be the bundle chosen with an initial money income of M_1 and x' the bundle chosen when money income rises to M_2 . If $x' \ll x^*$ i.e. if the demand for all goods is reduced, then x' must cost less than x^* since prices are held constant. x' was therefore available when x^* was chosen. But when x' was chosen x^* was still attainable (since money income had increased). The consumer therefore preferred x^* over x' with a money income of M_1 and x' over x^* with money income $M_2 > M_1$. He is therefore inconsistent: his behaviour violates the *transitivity assumption* of section A, and our model would have to be rejected.

If we now turn to the effects of changes in prices on the consumer's demands, Fig. 3.11 shows the implications of a fall in the price of x_1 with money income held constant. B_1 is the initial budget line, x^* the initial optimal bundle. A fall in p_1 , say from p_1 to p'_1 causes the budget line to shift to B_2 . x' is the optimal bundle on B_2 , x'' the optimal bundle on B_3 , which results from a further fall in p_1 from p'_1 to p''_1 . The *price consumption curve* (PCC) is traced out as the set of optimal bundles as p_1 varies. In this case the demand for both goods increases as p_1 falls. However, with different preferences the optimal bundle might have been x^0 or x^+ on B_2 . If x^0 was the optimal bundle with $p_1 = p'_1$ then x_1 would be a *Giffen good*, the demand for which falls as its price falls. We conclude that the demand for a good may fall, rise or remain unchanged as a result of a change in a price facing the consumer. Once again the model yields no definite (refutable) prediction about the effect

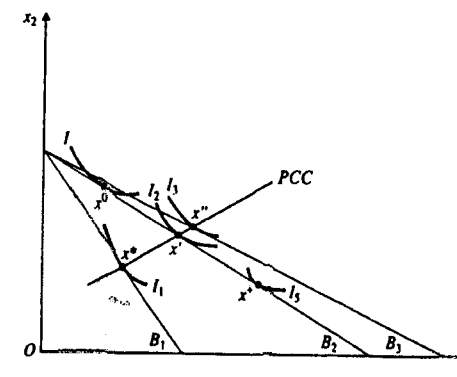


Fig. 3.11

on a *single* endogenous variable (the demand for a good) of a change in *one* of the exogenous variables (in this case a price). It is again possible, however, to predict (by reasoning similar to that employed in the case of a change in M) that a fall in price will not lead to a reduction in demand for *all* goods, and the reader should supply the argument.

Income and substitution effects

The analysis of the effect of price changes on the consumer's demands (optimal choices) has suggested that demand for a good may increase, decrease or remain unchanged, when its price rises; in other words anything may happen. We will now examine the effect of a change in the price of good 1 in more detail in order to see if it is possible to make more definite (refutable) predictions. We proceed by making a conceptual experiment. All we can ever actually observe is the change in quantity demanded following a price change. However, in order to say something more interesting about price-responses than we have been able to do so far, we carry out a hypothetical analysis which decomposes the overall demand change into two components. We then use this decomposition to say something more definite about consumer behaviour.

In Fig. 3.12, it can be seen that the fall in price of good 1 does two things:

- it reduces the expenditure required to achieve the initial utility level I_1 , allowing the higher utility level I_2 to be achieved with the same expenditure. Following J. R. Hicks, we then say that there has been an increase in the consumer's real income;
- it changes the relative prices facing the consumer.

In Fig. 3.12 we accordingly break down the change in demand for x_1 into:

- the income effect*, which is the change resulting solely from the change in real income, with relative prices held constant; and
- the own substitution effect*, which results solely from the change in p_1 with real income held constant.

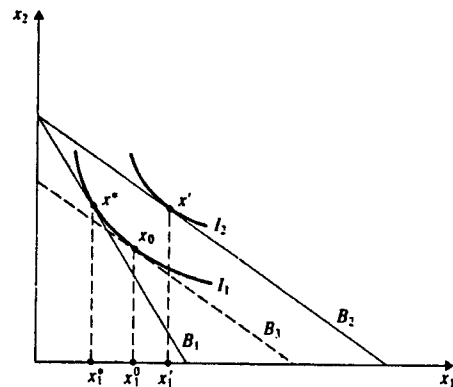


Fig. 3.12

x^* and x' are the optimal bundles before and after the fall in p_1 . B_1 and B_2 the corresponding budget lines. The *compensating variation* in money income is that change in M which will make the consumer just as well off after the price fall as he was before. In other words, there will be some reduction in M after the price fall which will 'cancel out' the real income gain and return the consumer to his initial indifference curve I_1 . The budget line is shifted inwards (reducing M) parallel with the post-price fall budget line B_2 until at B_3 it is just tangent to the original indifference curve I_1 . If the consumer were confronted with this budget line he would choose bundle x^0 . The difference between x^* and x^0 is due to a change in relative prices with real income (utility) held constant. The difference between x^0 and x' is due to the change in money income with relative prices held constant. x_1^* , x_1' and x_1^0 are the amounts of x_1 contained in the bundles x^* , x' , x^0 and

- $x_1^0 - x_1^*$ is the own substitution effect
- $x_1' - x_1^0$ is the income effect
- $(x_1^0 - x_1^*) + (x_1' - x_1^0) = x_1' - x_1^*$ is the total price effect.

The purpose of carrying out this experiment in hypothetical compensation is to isolate the fact that the own substitution effect will always be positive in the case of a price fall and negative for a price rise. The absolute value of the slope of the indifference curve declines from left to right, i.e. as more x_1 and less x_2 is consumed the curve flattens. The fall in p_1 flattens the slope of the budget line, and hence the budget line B_3 must be tangent with I_1 to the right of x^* , i.e. at a bundle containing more x_1 .

The income effect happens also to be positive in this particular case: x_1 is a normal good. If x_1 is inferior then the income effect is negative, x' contains less x_1 than x^0 and the price effect is smaller than the substitution effect. In Fig. 3.13(a) the income effect partially offsets the substitution effect but the price effect is still positive: a fall in p_1 leads to a rise in the demand for x_1 . In Fig. 3.13(b) the negative income effect more than offsets the positive substitution effect and x_1 is a Giffen good. Hence inferiority is a necessary, but not sufficient, condition for a good to be a Giffen good.

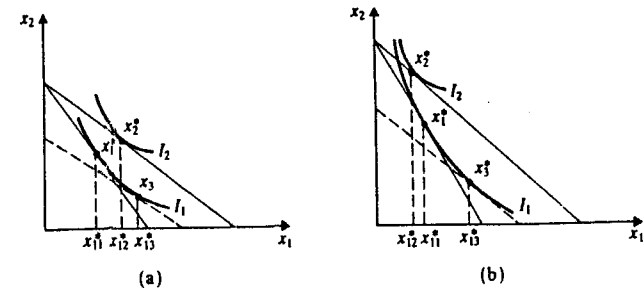


Fig. 3.13

This decomposition of the price effect has generated two further predictions:

1. *A normal good cannot be a Giffen good.* Hence, if we observe that a consumer increases his demand for a good when his money income rises (other things including prices being held constant), we would predict that if its price should fall, he will want to buy more of it. If we observe that he reduces his demand for the good when its price falls (and all other prices are constant and his money income is reasonably close to its original level), then the optimizing model of consumer behaviour has yielded a false prediction.
2. *The own substitution effect is always of opposite sign to the price change.*

The above decomposition of the price effect into an income and substitution effect is based on the definition, made by J. R. Hicks, of unchanged *real income* as an unchanged *utility level*. E. Slutsky suggested an alternative definition of a constant real income as the ability to purchase the bundle of goods bought before the price change. This *constant purchasing power* definition has the advantage that it does not require detailed knowledge of the consumer's indifference map.

Figure 3.14 reproduces Fig. 3.11 with some additions to show the relationship between the Hicks and Slutsky definitions of a constant real income. The budget line B_4 just enables the consumer to buy x^* , the initially optimal bundle, at the lower price of p_1 . Confronted with this budget line, the consumer actually chooses x^+ . The price effect has been decomposed into an income effect ($x_1' - x_1^*$) and an own substitution effect ($x_1^+ - x_1^*$). The income effect will again be positive, negative or zero depending on the form of the indifference map. The substitution effect will, as in the Hicksian case, always lead to a rise in demand for a good whose price has fallen. x^+ cannot lie to the left of x^* on B_4 because this would mean that the consumer is now choosing x^+ when x^* is still available, having previously rejected x^+ in favour of x^* . The transitivity assumption would be violated by such behaviour. The Slutsky definition yields a prediction (the sign of the substitution effect) which can be tested without specific knowledge of the consumer's indifference curves to 'cancel out' the income effect.

Our consideration of the comparative static properties of the model has shown that it does not yield refutable predictions about the overall change in demand for individual

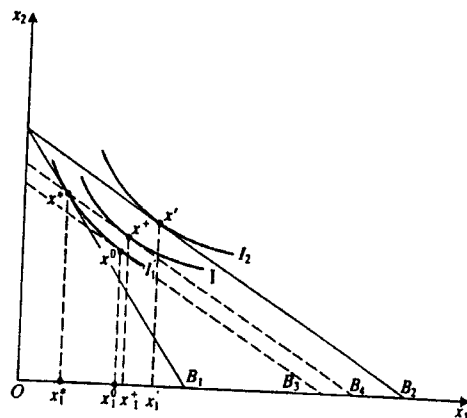


Fig. 3.14

goods induced by *ceteris paribus* changes in a price or money income. In other words

$$\frac{\partial x_i^*}{\partial p_j} = \frac{\partial D_i}{\partial p_j} \neq 0 \quad i, j = 1, 2, \dots, n$$

and

$$\frac{\partial x_i^*}{\partial M} = \frac{\partial D_i}{\partial M} \neq 0 \quad i, j = 1, 2, \dots, n$$

for every good and price. Only by considering the effect of changes in p_j or M on *all* goods, or by considering the effect of changes in p_j and M on a single good or by making more specific assumptions about the consumer's preferences can definite predictions be generated.

Consider, however, the consequences of equal proportionate changes in all prices and M . Suppose M increases to kM ($k > 1$) and prices to kp_1 and kp_2 . The slope of the budget line will be unaffected. The intercept on the x_1 axis is M/p_1 before the changes in M and prices and $kM/kp_1 = M/p_1$ after the change. Similarly for the intercept on the x_2 axis. Hence the equal proportionate changes in M and all prices alter neither the slope nor the intercepts on the budget line and so the feasible set is unaltered. If the feasible set is unchanged then so is the optimal bundle.

The model therefore predicts that the consumer will not suffer from *money illusion*; he will not alter his behaviour if his purchasing power and relative prices are constant, irrespective of the absolute level of prices and money income. More formally the demand function D_i for every commodity is *homogeneous of degree zero* in prices and money income, since we have:

$$x_i^* = D_i(kp, kM) = k^0 D_i(p, M) = D_i(p, M)$$

[D.2]

Demand curves

We complete this section on comparative statics by deriving the demand curve from the utility maximization model. The individual's *demand curve* for a good shows how his

desired or planned purchases of it vary as its price varies, other prices and income being held constant. As we have seen, a distinction can be drawn between constant *real* and constant *money* income and there are also two possible definitions of constant real income. Figure 3.15 shows the derivation of three demand curves, corresponding to the different assumptions about what is held constant, from the consumer's indifference map. The upper part of Fig. 3.15 is Fig. 3.14 with two additions. The *PCC* (price consumption curve) shows the bundles chosen as p_1 varies with M constant (i.e. as the budget line pivots through x_2^0). The constant purchasing power consumption curve (*CP*) shows the bundles chosen as p_1 varies, with the consumer's money income varying so as just to enable the consumer to purchase the original bundle x^* (i.e. the Slutsky definition of constant real income is adopted and so the budget line pivots through x^*). The indifference curve I_1 shows how consumption varies as p_1 varies, with M varying to keep the consumer's utility level constant (i.e. the Hicks definition of constant real income is adopted and so the budget line slides round I_1). These three curves therefore show the change in the demand for x_1 (and x_2) as p_1 changes, with income (variously defined) and p_2 held constant. The lower half of the figure uses the information contained in the three curves to plot demand curves for x_1 .

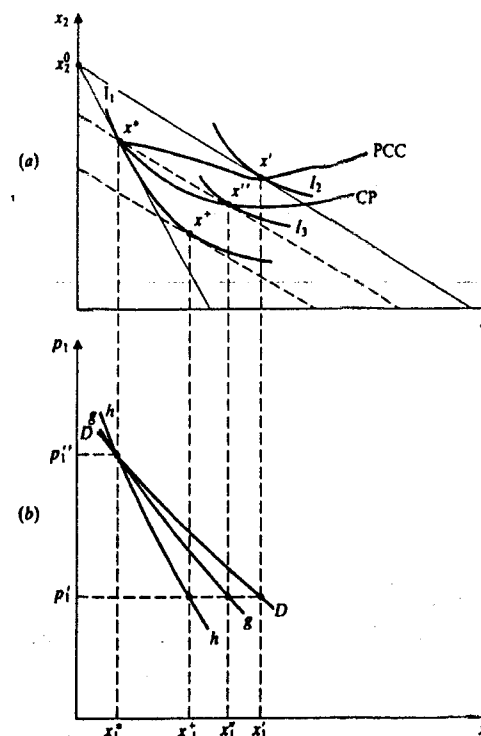


Fig. 3.15

pass through \bar{x} , and its slope will be unaffected by such price changes and hence its position is unchanged. Only changes in *relative* prices or in the initial endowments will alter the consumer's feasible set and therefore the consumer's demand or supply of a commodity. In the terminology of part D her demand functions will again be *homogeneous of degree zero in prices*.

It is clear from the way in which the budget constraint [E.1] was written that the consumer's optimization problem in this case is formally identical with that considered previously, so we will not dwell on the equilibrium conditions and the possibility of corner solutions. We restrict ourselves to examining the comparative static properties of tangency solutions and the derivation of supply and demand curves.

In Fig. 3.16, $x^* = (x_1^*, x_2^*)$ is the optimal consumption bundle on B , where the indifference curve I_1 is tangent to the budget line. Since $x_1^* > \bar{x}_1$ and $\bar{x}_2 > x_2^*$ the consumer is maximizing utility by selling commodity 2, which gives her receipts of $p_2(\bar{x}_2 - x_2^*)$, and buying commodity 1 at a cost of $p_1(x_1^* - \bar{x}_1)$.

Increases in p_1 relative to p_2 will make the budget line pivot clockwise about \bar{x} and the optimal bundle will vary as p_1/p_2 changes, as the upper half of Fig. 3.17 illustrates. With the budget line at B_2 the optimal bundle is the endowed bundle \bar{x} , and the consumer does not trade at all on the market. A further increase in p_1/p_2 will shift the budget line to B_3 where the optimal bundle is x' and the consumer is now selling commodity 1 and buying commodity 2.

The line FF in Fig. 3.17 is the locus of optimal bundles traced out as p_1/p_2 varies with \bar{x} fixed and is called the *offer curve*, since it shows the amounts (positive or negative) of the two goods which the consumer offers on the market at different relative prices. The consumer's *consumption demand curve* DD in the lower half of Fig. 3.17, which plots the consumption of x_1 as a function of p_1/p_2 , is derived from the offer curve. As p_1 increases relative to p_2 the consumer reduces consumption of commodity 1, from x_1^* to \bar{x}_1 and then to x'_1 as she moves along FF from x^* to \bar{x} and x' . $(p_1/p_2)_1$, $(p_1/p_2)_2$ and $(p_1/p_2)_3$ are the price ratios at which x^* , \bar{x} and x' are chosen.

The $\hat{D}\hat{D}$ curve in part (b) of Fig. 3.17 is the consumer's *net demand curve* and plots the net demand $\hat{x}_1 = x_1 - \bar{x}_1$, the amount of commodity 1 that she buys or sells on the market,

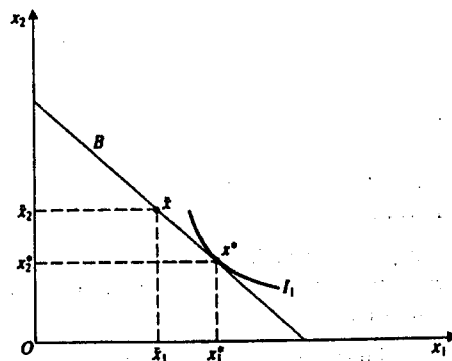


Fig. 3.16

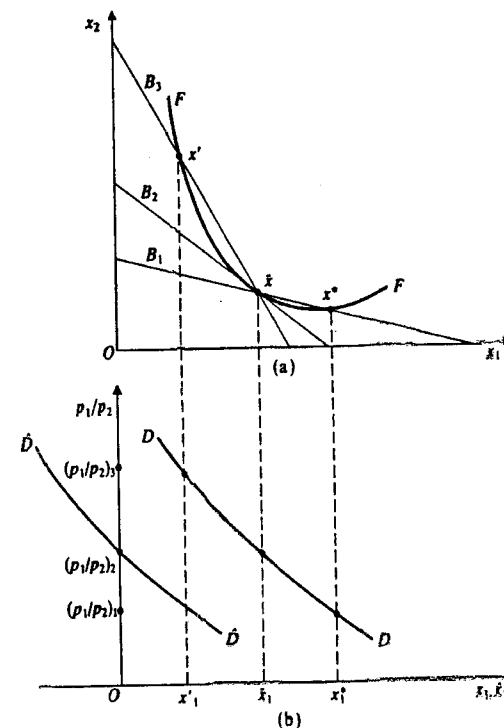


Fig. 3.17

against (p_1/p_2) . It is derived by taking the horizontal distance between the DD curve and a vertical line through $x_1 = \bar{x}_1$ at each price ratio. Notice that when $(p_1/p_2) > (p_1/p_2)_2$ the consumer's net demand is negative: she supplies commodity 1 to the market.

In the illustrations above the effect of a fall in the relative price of a commodity was to increase the consumer's demand for it. However, with a different indifference map the DD and $\hat{D}\hat{D}$ curves could have been positively sloped, indicating that rises in the relative price of commodity 1 reduce the amount of the commodity supplied to the market. Hence a *ceteris paribus* change in a single price may increase, decrease or leave unchanged the individual's consumption of any single commodity. Similarly, a *ceteris paribus* change in the initial endowment may increase, decrease or leave unchanged the consumption of commodity i . The comparative static properties of this model are so similar to those of the model in section C that we leave their derivation to the exercises at the end of the section.

The consumption decision in terms of net demands

The consumer's optimization problem studied earlier had the levels of consumption (x_1, \dots, x_n) as choice variables, but it is possible to formulate the problem with the consumer's net demands $(\hat{x}_1, \dots, \hat{x}_n)$ as the choice variables. Since this particular approach

constant money income demand curve DD shows the effect of changes in p_1 with M (and p_2) held constant. It plots the information contained in the price consumption curve. For example a fall in p_1 from p_1'' to p_1' with M constant causes the consumer to shift from bundle x^* to x' and his demand for x_1 to rise from x_1^* to x_1' .

The constant purchasing power demand curve gg corresponds to the CP curve. The fall in p_1 from p_1'' to p_1' with purchasing power constant causes the consumer to shift from x^* to x'' and his demand to increase from x_1^* to x_1'' .

The Hicksian constant utility demand curve hh is derived from the indifference curve I_1 . The fall in p_1 with utility constant at its initial level $u(x^*)$ causes the consumer to shift from x^* to x^+ , and his demand to increase from x_1^* to x_1^+ .

The constant money income demand curve plots the whole price effect and the other two curves plot only the two versions of the substitution effect. Hence the constant utility and purchasing power demand curves will be steeper than the constant money income demand curve when x_1 is a normal good, because they do not plot the income effect of the price change. When x_1 is inferior the relative steepness of the various demand curves is reversed. This analysis is taken further in Chapter 4, with the help of duality theory.

Exercise 3D

1. Derive the demand curve of the consumer of Question 3, Exercise 3B, for garbage disposal. Decompose the effects of a price change into income and substitution effects.
2. Examine the responses of an electricity consumer to changes in the connection charge and prices of electricity.
3. Examine the income and substitution effects in the cases given in Question 4, Exercise 3A.
4. Explain the difference between the Hicks and Slutsky definitions of real income, and apply this to explain why, in Fig. 3.15, the demand curve hh is steeper than the demand curve gg .
5. Why do we decompose the price effect into income and substitution effects?
- 6.* Examine the properties of the demand functions of a consumer with the following utility functions:
 - (a) $u(x) = x_1^{\alpha_1} \cdot x_2^{\alpha_2} \dots x_n^{\alpha_n}$ ($\alpha_i > 0$, all i ; $\sum \alpha_i = 1$) (Cobb-Douglas)
 - (b) $u(x) = (x_1 - k_1)^{\alpha_1} (x_2 - k_2)^{\alpha_2} \dots (x_n - k_n)^{\alpha_n}$ ($\alpha_i > 0$, all i ; $\sum \alpha_i = 1$; $k_i > 0$, all i) (Stone-Geary)
 - (c) $u(x) = \sum_i f_i(x_i)$ $f_i' > 0$, (additive separable)

What interpretation can be given to the k_i in case (b)?

E. Offer curves and net demand curves

We now consider the case of a consumer who has preferences satisfying the assumptions of section A, and is endowed, not with a given money income, but with fixed amounts of commodities which she can consume or sell on the market in order to finance purchases of other commodities. The feasible set is defined by the non-negativity requirements on consumption and by the constraint that the market value of the bundle consumed cannot exceed the market value of the consumer's initial endowments. Her budget constraint is therefore:

$$\sum p_i x_i \leq \sum p_i \bar{x}_i = W \quad i = 1, 2, \dots, n \quad [E.1]$$

where \bar{x}_i is her initial endowment of good i . $\sum p_i \bar{x}_i = W$ is the market value of the initial endowment, or the proceeds which could be obtained if the consumer sold all her initial endowments at the ruling market prices. Since W is a stock of value owned by the individual, we could think of it as her wealth.

If x_i (the amount of commodity i consumed) exceeds \bar{x}_i , i.e. if:

- (a) $\hat{x}_i = x_i - \bar{x}_i > 0$, then the consumer buys commodity i ; and if
- (b) $\hat{x}_i = x_i - \bar{x}_i < 0$, then she sells the commodity

where \hat{x}_i is defined as the net demand for commodity i . This suggests that we can re-write the budget constraint as:

$$\sum p_i \hat{x}_i = \sum p_i (x_i - \bar{x}_i) \leq 0 \quad i = 1, 2, \dots, n \quad [E.2]$$

which can be interpreted to mean that the sum of her expenditures on the quantities of goods she buys (which will be a positive component of the overall sum) cannot exceed the sum of the proceeds from the quantities of goods she sells (a negative component of the overall sum).

In the two-good case shown in Fig. 3.16 the budget line B is defined by

$$p_1 x_1 + p_2 x_2 = p_1 \bar{x}_1 + p_2 \bar{x}_2 = W \quad \text{or} \quad p_1 (x_1 - \bar{x}_1) + p_2 (x_2 - \bar{x}_2) = 0 \quad [E.3]$$

and has a slope of $-(p_1/p_2)$. B must pass through $\bar{x} = (\bar{x}_1, \bar{x}_2)$, the endowed bundle, since whatever prices she faces the consumer will always have the possibility of consuming her endowment, i.e. neither buying nor selling on the market. The feasible set is similar in shape to the case of the consumer endowed with a fixed money income M , but there are several significant differences as regards the effect of changes in prices:

- (a) the market value of the endowment $W = \sum p_i \bar{x}_i$ will increase or decrease as the price of a commodity increases or decreases.
- (b) Since the consumer is always able to consume her initial endowment vector \bar{x} , a change in a single price will cause the budget line to pivot through \bar{x} , rather than through an intercept on one of the axes.
- (c) An equal proportionate change in all prices leaves the budget line unaffected, though the value of the endowments varies in the same proportion. The budget line must still

The *Marshallian constant money income demand curve* DD shows the effect of changes in p_1 with M (and p_2) held constant. It plots the information contained in the price consumption curve. For example a fall in p_1 from p_1' to p_1'' with M constant causes the consumer to shift from bundle x^* to x' and his demand for x_1 to rise from x_1^* to x_1' .

The *constant purchasing power demand curve* gg corresponds to the CP curve. The fall in p_1 from p_1' to p_1'' with purchasing power constant causes the consumer to shift from x^* to x'' and his demand to increase from x_1^* to x_1'' .

The *Hicksian constant utility demand curve* hh is derived from the indifference curve I_1 . The fall in p_1 with utility constant at its initial level $u(x^*)$ causes the consumer to shift from x^* to x^+ , and his demand to increase from x_1^* to x_1^+ .

The constant money income demand curve plots the whole price effect and the other two curves plot only the two versions of the substitution effect. Hence the constant utility and purchasing power demand curves will be steeper than the constant money income demand curve when x_1 is a normal good, because they do not plot the income effect of the price change. When x_1 is inferior the relative steepness of the various demand curves is reversed. This analysis is taken further in Chapter 4, with the help of duality theory.

Exercise 3D

- Derive the demand curve of the consumer of Question 3, Exercise 3B, for garbage disposal. Decompose the effects of a price change into income and substitution effects.
 - Examine the responses of an electricity consumer to changes in the connection charge and prices of electricity.
 - Examine the income and substitution effects in the cases given in Question 4, Exercise 3A.
 - Explain the difference between the Hicks and Slutsky definitions of real income, and apply this to explain why, in Fig. 3.15, the demand curve hh is steeper than the demand curve gg .
 - Why do we decompose the price effect into income and substitution effects?
 - Examine the properties of the demand functions of a consumer with the following utility functions:
 - $u(x) = x_1^{\alpha_1} \cdot x_2^{\alpha_2} \dots x_n^{\alpha_n}$ ($\alpha_i > 0$, all i ; $\sum \alpha_i = 1$) (Cobb-Douglas)
 - $u(x) = (x_1 - k_1)^{\alpha_1} (x_2 - k_2)^{\alpha_2} \dots (x_n - k_n)^{\alpha_n}$ ($\alpha_i > 0$, all i ; $\sum \alpha_i = 1$; $k_i > 0$, all i) (Stone-Geary)
 - $u(x) = \sum f_i(x_i)$ $f_i' > 0$, (additive separable)
- What interpretation can be given to the k_i in case (b)?

E. Offer curves and net demand curves

We now consider the case of a consumer who has preferences satisfying the assumptions of section A, and is endowed, not with a given money income, but with fixed amounts of commodities which she can consume or sell on the market in order to finance purchases of other commodities. The feasible set is defined by the non-negativity requirements on consumption and by the constraint that the market value of the bundle consumed cannot exceed the market value of the consumer's initial endowments. Her budget constraint is therefore:

$$\sum p_i \bar{x}_i \leq \sum p_i \bar{x}_i = W \quad i = 1, 2, \dots, n \quad [E.1]$$

where \bar{x}_i is her *initial endowment of good i* , $\sum p_i \bar{x}_i = W$ is the market value of the initial endowment, or the proceeds which could be obtained if the consumer sold all her initial endowments at the ruling market prices. Since W is a stock of value owned by the individual, we could think of it as her wealth.

If x_i (the amount of commodity i consumed) exceeds \bar{x}_i , i.e. if:

- $\hat{x}_i = x_i - \bar{x}_i > 0$, then the consumer buys commodity i ; and if
- $\hat{x}_i = x_i - \bar{x}_i < 0$, then she sells the commodity

where \hat{x}_i is defined as the *net demand* for commodity i . This suggests that we can re-write the budget constraint as:

$$\sum p_i \hat{x}_i = \sum p_i (x_i - \bar{x}_i) \leq 0 \quad i = 1, 2, \dots, n \quad [E.2]$$

which can be interpreted to mean that the sum of her expenditures on the quantities of goods she buys (which will be a positive component of the overall sum) cannot exceed the sum of the proceeds from the quantities of goods she sells (a negative component of the overall sum).

In the two-good case shown in Fig. 3.16 the budget line B is defined by

$$p_1 x_1 + p_2 x_2 = p_1 \bar{x}_1 + p_2 \bar{x}_2 = W \quad \text{or} \quad p_1 (x_1 - \bar{x}_1) + p_2 (x_2 - \bar{x}_2) = 0 \quad [E.3]$$

and has a slope of $-(p_1/p_2)$. B must pass through $\bar{x} = (\bar{x}_1, \bar{x}_2)$, the endowed bundle, since whatever prices she faces the consumer will always have the possibility of consuming her endowment, i.e. neither buying nor selling on the market. The feasible set is similar in shape to the case of the consumer endowed with a fixed money income M , but there are several significant differences as regards the effect of changes in prices:

- the market value of the endowment $W = \sum p_i \bar{x}_i$ will increase or decrease as the price of a commodity increases or decreases.
- Since the consumer is always able to consume her initial endowment vector \bar{x} , a change in a single price will cause the budget line to pivot through \bar{x} , rather than through an intercept on one of the axes.
- An equal proportionate change in all prices leaves the budget line unaffected, though the value of the endowments varies in the same proportion. The budget line must still

pass through \bar{x} , and its slope will be unaffected by such price changes and hence its position is unchanged. Only changes in *relative* prices or in the initial endowments will alter the consumer's feasible set and therefore the consumer's demand or supply of a commodity. In the terminology of part D her demand functions will again be *homogeneous of degree zero in prices*.

It is clear from the way in which the budget constraint [E.1] was written that the consumer's optimization problem in this case is formally identical with that considered previously, so we will not dwell on the equilibrium conditions and the possibility of corner solutions. We restrict ourselves to examining the comparative static properties of tangency solutions and the derivation of supply and demand curves.

In Fig. 3.16, $x^* = (x_1^*, x_2^*)$ is the optimal consumption bundle on B , where the indifference curve I_1 is tangent to the budget line. Since $x_1^* > \bar{x}_1$ and $\bar{x}_2 > x_2^*$ the consumer is maximizing utility by selling commodity 2, which gives her receipts of $p_2(\bar{x}_2 - x_2^*)$, and buying commodity 1 at a cost of $p_1(x_1^* - \bar{x}_1)$.

Increases in p_1 relative to p_2 will make the budget line pivot clockwise about \bar{x} and the optimal bundle will vary as p_1/p_2 changes, as the upper half of Fig. 3.17 illustrates. With the budget line at B_2 the optimal bundle is the endowed bundle \bar{x} , and the consumer does not trade at all on the market. A further increase in p_1/p_2 will shift the budget line to B_3 where the optimal bundle is x' and the consumer is now selling commodity 1 and buying commodity 2.

The line FF in Fig. 3.17 is the locus of optimal bundles traced out as p_1/p_2 varies with \bar{x} fixed and is called the *offer curve*, since it shows the amounts (positive or negative) of the two goods which the consumer offers on the market at different relative prices. The consumer's *consumption demand curve* DD in the lower half of Fig. 3.17, which plots the consumption of x_1 as a function of p_1/p_2 , is derived from the offer curve. As p_1 increases relative to p_2 the consumer reduces consumption of commodity 1, from x_1^* to \bar{x}_1 and then to x'_1 as she moves along FF from x^* to \bar{x} and x' . $(p_1/p_2)_1$, $(p_1/p_2)_2$ and $(p_1/p_2)_3$ are the price ratios at which x^* , \bar{x} and x' are chosen.

The $\hat{D}\hat{D}$ curve in part (b) of Fig. 3.17 is the consumer's *net demand curve* and plots the net demand $\hat{x}_1 = x_1 - \bar{x}_1$, the amount of commodity 1 that she buys or sells on the market,

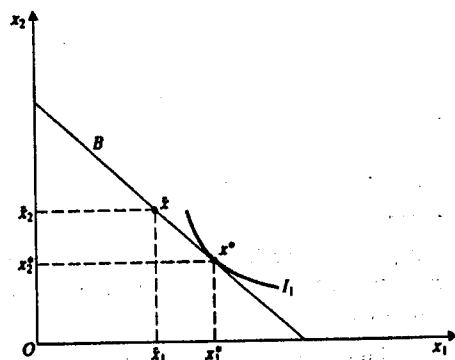


Fig. 3.16

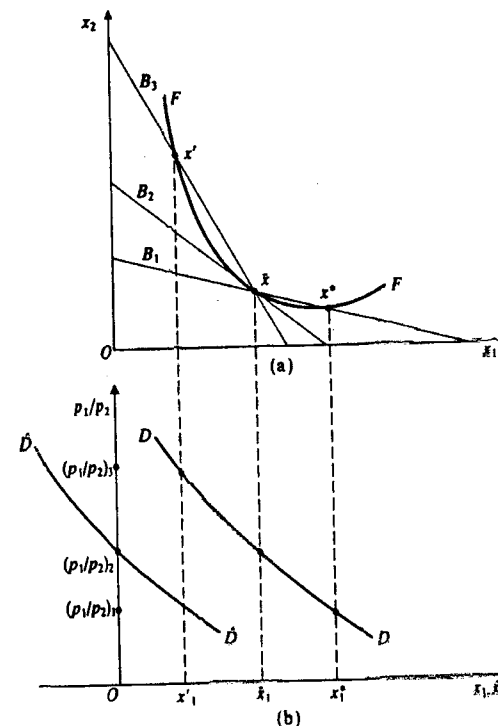


Fig. 3.17

against (p_1/p_2) . It is derived by taking the horizontal distance between the DD curve and a vertical line through $x_1 = \bar{x}_1$ at each price ratio. Notice that when $(p_1/p_2) > (p_1/p_2)_2$ the consumer's net demand is negative: she *supplies* commodity 1 to the market.

In the illustrations above the effect of a fall in the relative price of a commodity was to increase the consumer's demand for it. However, with a different indifference map the DD and $\hat{D}\hat{D}$ curves could have been positively sloped, indicating that rises in the relative price of commodity 1 reduce the amount of the commodity supplied to the market. Hence a *ceteris paribus* change in a single price may increase, decrease or leave unchanged the individual's consumption of any single commodity. Similarly, a *ceteris paribus* change in the initial endowment may increase, decrease or leave unchanged the consumption of commodity i . The comparative static properties of this model are so similar to those of the model in section C that we leave their derivation to the exercises at the end of the section.

The consumption decision in terms of net demands

The consumer's optimization problem studied earlier had the levels of consumption (x_1, \dots, x_n) as choice variables, but it is possible to formulate the problem with the consumer's net demands $(\hat{x}_1, \dots, \hat{x}_n)$ as the choice variables. Since this particular approach

will be used in Chapter 16 on general equilibrium, it is useful to show here that it is equivalent to the model of section C.

The consumer's utility function $u(x)$ can be rewritten as a function of the net demands \hat{x}_i since from the definitions $\hat{x}_i \equiv x_i - \bar{x}_i$ we have $x_i \equiv \hat{x}_i + \bar{x}_i$ and so

$$u(x_1, \dots, x_n) = u(\hat{x}_1 + \bar{x}_1, \dots, \hat{x}_n + \bar{x}_n) \quad [E.4]$$

Since the initial endowments \bar{x}_i are constants, u varies only as the \hat{x}_i vary:

$$u(\hat{x}_1 + \bar{x}_1, \dots, \hat{x}_n + \bar{x}_n) = \hat{u}(\hat{x}_1, \dots, \hat{x}_n) \quad [E.5]$$

and

$$\frac{\partial \hat{u}}{\partial \hat{x}_i} = \frac{\partial u}{\partial x_i} = \frac{\partial u}{\partial \hat{x}_i} \frac{\partial \hat{x}_i}{\partial x_i} = \frac{\partial u}{\partial x_i} \quad [E.6]$$

\hat{u} will have all the properties possessed by u such as continuity, quasi-concavity, etc.

The feasible set can also be rewritten in terms of the \hat{x}_i , as was shown in [E.2].

The non-negativity constraints on the x_i are replaced by

$$x_i = \hat{x}_i + \bar{x}_i \geq 0 \quad [E.7]$$

or

$$\hat{x}_i \geq -\bar{x}_i \quad (j = 1, \dots, n)$$

i.e. the supply of a good cannot exceed the endowment of that good. The consumer's optimization problem can now be written in terms of the net demands as (compare [C.1]):

$$\begin{aligned} \max \hat{u}(\hat{x}_1, \dots, \hat{x}_n) \\ \text{s.t. } \sum p_i \hat{x}_i &\leq 0 \\ \hat{x}_i &\geq -\bar{x}_i \quad (i = 1, \dots, n) \end{aligned} \quad [E.8]$$

Proceeding, as in section C, to assume that the direct constraints [E.7] do not bind at the solution, the Lagrange function of the problem may be written

$$L = \hat{u}(\hat{x}_1, \dots, \hat{x}_n) - \lambda \sum p_i \hat{x}_i$$

First-order conditions are

$$\frac{\partial L}{\partial \hat{x}_i} = \hat{u}_i - \lambda p_i = 0 \quad (i = 1, \dots, n) \quad [E.9]$$

$$\frac{\partial L}{\partial \lambda} = -\sum p_i \hat{x}_i = 0 \quad [E.10]$$

and from [E.6] we see that [E.9] is identical to [C.7], so that we would be able to derive exactly the same equilibrium conditions and comparative static properties as in section C.

This reformulation of the problem in terms of net demands rather than consumption bundles is, in terms of the diagrammatic analysis of Fig. 3.17, equivalent to shifting the origin to \bar{x} so that $\hat{x}_1 = x_1 - \bar{x}_1$ and $\hat{x}_2 = x_2 - \bar{x}_2$ are measured along the axes. The budget line now passes through the new origin and the consumer's indifference map is unaffected.

The reader should redraw Fig. 3.15 and the upper part of 3.16 in this way to confirm that nothing of substance is affected by the relabelling.

Exercise 3E

- 1.* Show that all the predictions of sections 3C, 3D hold when the analysis is recast in terms of net demands.
2. How would you interpret the slope of the consumer's offer curve?
3. Explain why, at every point on the offer curve, there is tangency between an indifference curve and the budget line generating that point.
- 4.* Discuss the relevance of the model examined in this section to:
 - (a) a market in stocks and shares;
 - (b) a market in new and secondhand cars;
 - (c) the case where x_1 is bread and x_2 is leisure time, with $\bar{x}_1 = 0$ and $\bar{x}_2 = 24$ hours per day.

Appendix 1: The lexicographic ordering*

The significance of the lexicographic ordering is that it shows the need for the continuity assumption, if we wish to work with a numerical representation of the consumer's preference ordering, i.e. with a utility function. The lexicographic ordering can be shown to satisfy the first four assumptions set out in section A but to be incapable of being represented by a utility function. On the other hand, it can be shown to give rise to well-defined demand functions, which implies that the continuity assumption is certainly not necessary for the existence of these.

The ordering takes the following form. Suppose consumption bundles consist only of two goods, i.e. $x = (x_1, x_2)$. Then the consumer's preferences are such that, given two bundles $x' = (x'_1, x'_2)$ and $x'' = (x''_1, x''_2)$:

- (a) $x'_1 > x''_1$ implies $x' \succ x''$
- (b) $x'_1 = x''_1$ and $x'_2 > x''_2$ implies $x' \succ x''$

In words: the consumer always prefers a bundle with more of the first good in it, regardless of the quantity of the second good; only if the bundles contain the same amount of the first good does the quantity of the second matter. An illustration would be the case of a drunkard who would always prefer a combination of beer and bread with more beer in it to one with less, regardless of the amount of bread, but if the amounts of beer are the same, he will prefer the one with more bread. It is called a 'lexicographic ordering' because it is analogous to the way words are ordered in a dictionary: A always comes before B, but if two words begin with A then the second letter determines the order they are placed in. Here the goods play the role of letters.

The indifference sets corresponding to this ordering are found with the help of Fig. 3.18. Take the bundle $x' = (x'_1, x'_2)$, and ask: what points are preferred to it, and to what points is it preferred? The area B , including the points on the solid line above x' , must all be preferred to x' , since points to the right of x' contain more x_1 , while points along the solid line contain as much x_1 and more x_2 . The area W , including the points on the broken line below x' , must all be such that x' is preferred to them, since points to the left of x' contain less x_1 , while points on the broken line contain as much x_1 but less x_2 . But if all the points in B are preferred to x' , and x' is preferred to all the points in W , there can be no other points indifferent to x' , and so the indifference set for x' consists only of this single point. Since x' was chosen arbitrarily, this is true for every point in the space: each lies in an indifference set consisting only of itself.

The lexicographic ordering does not satisfy the assumption of continuity, since the indifference set is a point and not a continuous surface. If we reduce the amount of x_1 in the bundle, by however small an amount, we can find no amount of x_2 to compensate for the change (the drunkard cannot be bribed by any amount of bread to give up even a sip of beer). We now show that it is not possible to represent the lexicographic ordering by a utility function.

First, we know that if we divide the real line into non-empty, disjoint bounded intervals, the set of these intervals is *countable*. That is, we can put them into a one-to-one correspondence with the set of positive integers, $\{1, 2, 3, \dots\}$. On the other hand, the points on the real line itself or some interval of it, e.g. its positive half, are *not* countable. It follows that any argument which leads to the conclusion that the positive half of the real line is countable must be false. What we can show is that the assumption that a utility function exists for the lexicographic ordering does just that.

Suppose then that a utility function $u(x_1, x_2)$ exists, which gives a numerical representation of the lexicographic ordering. Refer to Fig. 3.18. Setting $x_1 = x'_1$, the function will take on the values $u(x'_1, x_2)$ for all $x_2 \geq 0$ along the vertical line in the figure. This

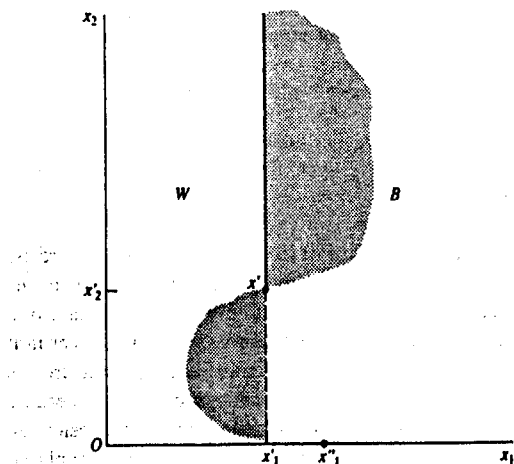


Fig. 3.18

set of values has a lower bound at $u(x'_1, 0)$, and it must have an upper bound because the u -values for any $x_1 = x'_1 > x'_1$ must be greater. Hence this set of values represents a non-empty bounded interval on the real line. Now choosing a value $x'_1 > x'_1$ we can in the same way associate with it a non-empty bounded interval of real numbers. Moreover, however close to x'_1 we choose x'_1 , it must always be the case that the intervals are disjoint, since $u(x'_1, x_2) > u(x'_1, x_2)$ for every x_2 , since $x'_1 > x'_1$. We can repeat this argument for every value of x_1 on the horizontal axis: to each corresponds a unique non-empty bounded interval on the real line. But this means we have put the positive half of the real line into one-to-one correspondence with a set of non-empty disjoint intervals of the real line, implying that the former is countable. This is false, and therefore so is the initial assumption that the utility function exists.

As we suggested earlier, the fact that the lexicographic ordering does not possess a numerical representation does not mean the consumer's choice problem cannot be solved nor even that a continuous demand function does not exist. Remember that continuity assumptions are usually sufficient rather than necessary. Thus consider our beer drinker with lexicographic preferences, an income of M and facing price p_1 for beer and price p_2 for bread. He will *always* spend his entire income on beer and nothing on bread. Hence his demand functions are

$$x_1 = M/p_1, \quad x_2 = 0$$

which are well-defined and continuous. His demand curve for beer is just a rectangular hyperbola in (x_1, p_1) space and his demand function for bread in (x_2, p_2) space is the vertical axis.

Exercise A1

1. Suppose that the consumer has lexicographic preferences, but must consume a minimum level of x_2 for subsistence. Show how this affects his demand functions.
2. Likewise, show how the analysis is affected by the assumption that the consumer reaches a satiation level for x_1 , but not for x_2 .
3. Generalize the statement of the lexicographic ordering to n goods. What would be the demand functions with and without subsistence and satiation levels of each good?
4. How plausible do you find the assumption that a consumer has a lexicographic preference ordering with respect to:
 - (a) each good taken separately;
 - (b) groups of goods, e.g. food, clothing, shelter, entertainment?

Appendix 2: Existence of a utility function*

The lexicographic ordering satisfies completeness, reflexivity, transitivity and non-satiation but no utility function can be constructed to represent it. We now show that adding the

assumption of continuity guarantees that a continuous, increasing utility function can be found to represent the preference ordering. We do this by actually constructing such a function.

The basic idea is illustrated in Fig. 3.19. Since the indifference curves are continuous, they intersect the 45° line as shown in the figure. Then, for any point such as x^0 , associate with it the real number $u(x^0)$, a coordinate of the point at which the indifference curve through x^0 cuts the 45° line (either coordinate would obviously do just as well). Then the u -values found in this way give us a numerical representation of this utility function: indifferent bundles have the same u -value, preferred bundles have higher u -values. We now put this more formally.

First, we need to state the continuity axiom in a more precise form. The better set of any point x^0 is $B(x^0) = \{x | x \succ x^0\}$, and the worse set is $W(x^0) = \{x | x^0 \succ x\}$. Then the continuity assumption is

For all bundles x^0 the sets $B(x^0)$ and $W(x^0)$ are closed.

Let us see first how the lexicographic ordering violates this assumption. Recall that a closed set is one which contains its boundary points, i.e. points having the property that every neighbourhood of them contains points which are, and points which are not, in the set. In Fig. 3.18, the boundary of $B(x')$ and $W(x')$ is the vertical line through x' . Points on the dotted line below x' are not in $B(x')$, points on the solid line above x' are not in $W(x')$, and so neither of these sets is closed. Loosely, this means that one can move from a point strictly preferred to x' , to a point strictly inferior to x' , without passing through a point indifferent to x' , however close to x' these points may be. In this loose sense therefore there is a 'jump' in the preference ordering. The continuity assumption removes this possibility.

We now construct the utility function. First, we show that associated with any x^0 is a unique number $u(x^0)$, then that this has the properties of a utility function, and finally that this function is continuous.

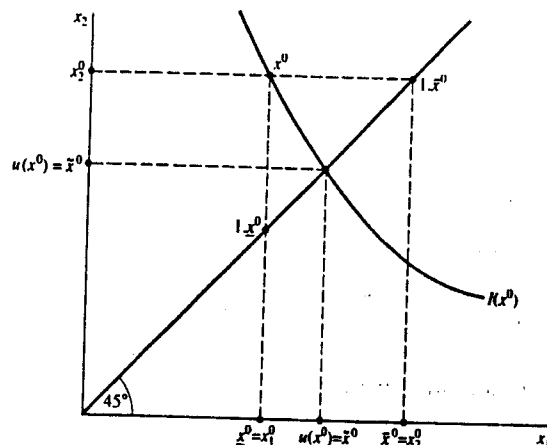


Fig. 3.19

Existence of $u(x^0)$: Given any bundle $x^0 = (x_1^0, \dots, x_n^0)$, choose the smallest and largest components, and denote them by \underline{x}^0 and \bar{x}^0 respectively. If 1 is the n -vector $(1, \dots, 1)$, then we have $1 \cdot \bar{x}^0 \succ x^0 \succ 1 \cdot \underline{x}^0$ (Fig. 3.19 illustrates).

If $\bar{x}^0 = \underline{x}^0$ then the proposition holds trivially so we ignore this case. Consider the non-empty interval of real numbers $[\underline{x}^0, \bar{x}^0]$. We claim that there exists a number $\bar{x}^0 \in [\underline{x}^0, \bar{x}^0]$ such that $1 \cdot \bar{x}^0 \sim x^0$. Suppose not. Then for every $x \in [\underline{x}^0, \bar{x}^0]$ we have either $1 \cdot x \succ x^0$ or $x^0 \succ 1 \cdot x$. Moreover, the transitivity and non-satiation assumptions imply that those x -values for which $1 \cdot x \succ x^0$ form a sub-interval of numbers strictly greater than the complementary sub-interval for which $x^0 \succ 1 \cdot x$. This implies that the former sub-interval has a lower bound, and so a greatest lower bound; likewise the latter sub-interval has a least upper bound. Moreover, these bounds must be the same. Denote this common bound by b . Thus we can partition $[\underline{x}^0, \bar{x}^0]$ into $[\underline{x}^0, b]$ and $(b, \bar{x}^0]$, or $[\underline{x}^0, b)$ and $[b, \bar{x}^0]$. In each case, $1 \cdot b$ is a boundary point of $W(x^0)$ and $B(x^0)$, and is either not contained in $W(x^0)$ or not contained in $B(x^0)$. But this contradicts the assumption that these sets are closed. So there does exist \bar{x}^0 such that $1 \cdot \bar{x}^0 \sim x^0$, and we take $u(x^0) = \bar{x}^0$. The non-satiation assumption implies that $u(x^0)$ is unique.

We now have to show that the $u(x)$ numbers constructed in this way satisfy the definition of a utility function, which is, for any two bundles x^0, x' ,

$$u(x^0) \geq u(x') \Leftrightarrow x^0 \succsim x'$$

Proof:

(a) $u(x^0) \geq u(x') \Rightarrow x^0 \succsim x'$. Suppose not, i.e. $u(x^0) \geq u(x')$ but $x' \succ x^0$. Then $1 \cdot u(x^0) \geq 1 \cdot u(x')$, where 1 is again an n -vector of ones. We then have by transitivity

$$1 \cdot u(x') \sim x' \succ x^0 \sim 1 \cdot u(x^0)$$

which by non-satiation gives the contradiction

$$u(x') > u(x^0)$$

(b) $x^0 \succsim x' \Rightarrow u(x^0) \geq u(x')$. Suppose not, i.e. $x^0 \succsim x'$ but $u(x') > u(x^0)$. Then $1 \cdot u(x') > 1 \cdot u(x^0)$ and the chain $x' \sim 1 \cdot u(x') > 1 \cdot u(x^0) \sim x^0$ gives the contradiction.

Finally, to prove that $u(x)$ is a continuous function it is convenient to take the following property of continuous functions (see, for example, K. Binmore, *Topological Ideas*, Cambridge, Cambridge University Press, 1981, ch. 16).

A function $u(x)$, $x \in \mathbb{R}_+^n$, is continuous on \mathbb{R}_+^n if and only if for each pair of subsets of function values, U_1 and U_2 , if U_1 and U_2 are separated then $u^{-1}(U_1)$ and $u^{-1}(U_2)$ are separated.

Two sets are separated if no point in one set is a boundary point of the other. Thus, for example, the pairs of sets $[0, 3/4]$ and $[1, 2]$, and $[0, 1)$ and $(1, 2]$ are separated, while the pair $[0, 1]$ and $(1, 2]$ are not.

Then, take any bundle x^0 and its corresponding utility value $u(x^0)$, and form the intervals $U_1 = [\underline{u}, u(x^0))$, $U_2 = (u(x^0), \bar{u}]$ where $\underline{u} < u(x^0)$ and $\bar{u} > u(x^0)$ are arbitrary. Then clearly U_1 and U_2 are separated. The set $u^{-1}(U_1) = \{x | u(x) \in U_1\}$ is a subset of the interior of $W(x^0)$, and the set $u^{-1}(U_2) = \{x | u(x) \in U_2\}$ is a subset of the interior of $B(x^0)$. Since these subsets lie on either side of $I(x^0)$, which belongs to neither of them, they are also separated. Since x^0 , \underline{u} and \bar{u} were arbitrary, the function $u(x)$ is continuous.

The discussion of the existence of a continuous, increasing utility function in this Appendix used a specific construction and also made use of the non-satiation assumption. It is possible to drop this assumption and still prove existence of a continuous utility function, but this requires mathematical methods outside the prerequisites for this book.

References and further reading

A general introduction to the theory of consumer behaviour and its applications is

H. A. J. Green. *Consumer Theory*, Macmillan, London, 1976.

Consumer preferences are considered in some detail in

P. Newman. *The Theory of Exchange*, Prentice-Hall, Englewood Cliffs, NJ, 1965, ch. 2,

and at a more advanced level in

G. Debreu. *Theory of Value*, John Wiley, New York, 1959, ch. 4.

An excellent account of the history of utility theory is

G. Stigler. 'The development of utility theory', *Journal of Political Economy*, 58, 1950, 307-27, 373-96.

CHAPTER 4

Consumer theory: duality

In the previous chapter we defined the consumer problem as that of choosing a vector x to solve the problem $\max u(x)$ s.t. $px = M$, where p is a price vector and M money income. From the solution we derived n Marshallian demand functions $x_i = D_i(p, M)$, $i = 1, \dots, n$, which express demands as functions of prices and money income. We observed that we cannot place restrictions on the signs of the partial derivatives of these functions: $\partial D_i / \partial M \gtrless 0$, $\partial D_i / \partial p_j \gtrless 0$, $i, j = 1, \dots, n$. In particular the demand for a good does not necessarily vary inversely with its own price. However, as a result of a diagrammatic analysis, we were able to say that this will be true of normal goods, or of inferior goods whose income effects are weaker than their substitution effects. We now put this analysis on a more rigorous and general basis. We also consider the problem, central to many applications of consumer theory, of deriving a money measure of the costs and benefits incurred by a consumer as a result of price changes. In doing so, we develop the methods and concepts of *duality theory*, an approach to the analysis of optimization problems which permits an elegant and concise derivation of comparative static results.

A. The expenditure function

The expenditure function is derived from the problem of minimizing the total expenditure necessary for the consumer to achieve a specified level of utility u :

$$\min_{x_1, \dots, x_n} \sum p_i x_i \text{ s.t. (i) } u(x_1, \dots, x_n) \geq u \quad [A.1]$$

$$(ii) \ x_i \geq 0 \quad (i = 1, \dots, n)$$

If all prices are positive the first constraint in [A.1] will be satisfied as an equality in the solution, since if $u(x) > u$ expenditure can be reduced without violating the constraint. If it is further assumed that all x_i are strictly positive in the solution, we can write the Lagrange function for the problem (with μ as the Lagrange multiplier) as

$$L = \sum p_i x_i + \mu [u - u(x_1, \dots, x_n)] \quad [A.2]$$

and the necessary conditions for a minimum of L , also the necessary conditions for a solution of [A.1], are

$$\frac{\partial L}{\partial x_i} = p_i - \mu u_i = 0 \quad (i = 1, \dots, n) \quad [\text{A.3}]$$

$$\frac{\partial L}{\partial \mu} = u - u(x_1, \dots, x_n) = 0 \quad [\text{A.4}]$$

The conditions on the x_i bear a striking resemblance to [C.9] in Chapter 3. Writing them as $p_i = \mu u_i$ and dividing the condition on x_i by the condition on x_j gives

$$\frac{p_i}{p_j} = \frac{u_i}{u_j} \quad [\text{A.5}]$$

which is identical with Chapter 3 [C.2]: the ratio of prices is equated to the marginal rate of substitution. This is not surprising as examination of the two-good case in Fig. 4.1 indicates. The indifference curve I_0 shows the combinations of x_1 and x_2 which give a utility level of u and the feasible set for the problem is all points on or above I_0 . The lines m_0, m_1, m_2 , are isoexpenditure lines similar to the budget lines of earlier diagrams. m_0 , for example, plots all bundles costing m_0 , i.e. satisfying the equation $p_1 x_1 + p_2 x_2 = m_0$. The problem is to find the point in the feasible set which is on the lowest isoexpenditure line. This will, in the tangency solution shown here, be where the indifference curve I_0 is tangent to the isoexpenditure line m_0 . The problem confronting the utility maximizing consumer is to move along her budget line until the highest indifference curve is reached. The expenditure minimizing problem is to move along the indifference curve until the lowest isoexpenditure line is reached.

The optimal x_i^* in problem [A.1] depend on the prices and the utility level u :

$$x_i^* = H_i(p_1, \dots, p_n, u) = H_i(p, u) \quad (i = 1, \dots, n) \quad [\text{A.6}]$$

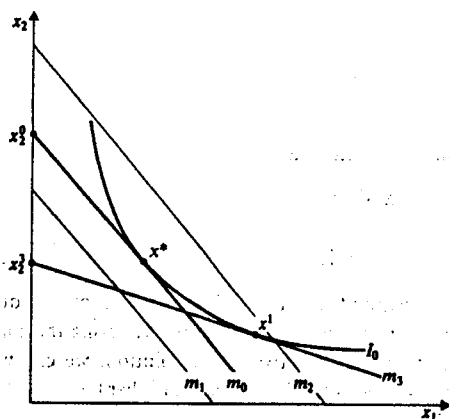


Fig. 4.1

and $H_i(p, u)$ is the Hicksian demand function for x_i . Substituting the optimal values of the x_i in $\sum p_i x_i$ gives

$$\sum p_i x_i^* = \sum p_i H_i(p, u) = m(p, u) \quad [\text{A.7}]$$

$m(p, u)$ is the expenditure function, showing the minimum level of expenditure necessary to achieve a given utility level as a function of prices and the required utility level.

The Hicksian demand function is also called the compensated demand function. In considering the effect of a change in price on demand with utility held constant (the partial derivative $\partial H_i / \partial p_i$, $i, j = 1, \dots, n$), we automatically make whatever changes in expenditure are required to compensate for the effects of the price change on real income or utility. This is illustrated in Fig. 4.1. Assume p_2 remains constant while p_1 falls to give a new family of isoexpenditure lines, with slopes corresponding to that of m_3 in the figure. x^1 is the new expenditure minimizing consumption bundle, and the change from x^* to x^1 is the effect of making the relative price change with m varying to keep u constant. The optimal expenditure line slides round the indifference curve from m_0 to m_3 as the optimal bundle changes from x^* to x^1 . The minimized total expenditure can be read off from the intercepts of m_0 and m_3 on the x_2 axis. The fall in p_1 lowers m from $p_2 x_2^0$ to $p_2 x_2^1$.

Provided the indifference curves are strictly convex to the origin the optimal x_i (and hence the expenditure function) vary smoothly and continuously with the prices of the goods. Hence the $H_i(p, u)$ functions have continuous derivatives with respect to the prices. The demand curve we derive from the Hicksian demand function was represented by curve hh in Fig. 3.15. The slope of this Hicksian or compensated demand curve, $\partial H_i / \partial p_i$, $i = 1, \dots, n$, is the substitution effect of the price change, since by definition $\partial H_i / \partial p_i$ is taken with u held constant.

We later consider in depth the properties of the Hicksian demand function. First, however, we analyse the expenditure function in some detail. The expenditure function gives the smallest expenditure, at a given price vector, that is required to achieve a particular 'standard of living' or utility level, and describes how that expenditure will change as prices or the required utility level change. The assumptions made in Chapter 3 concerning the nature of the consumer's preference ordering and indifference sets imply certain properties of the expenditure function:

(a) The expenditure function is concave in prices. Choose two price vectors p' and p'' , and k such that $0 \leq k \leq 1$. Define $\bar{p} = kp' + (1 - k)p''$. We have to prove that (recall the definition of concavity in section 2B):

$$m(\bar{p}, u) \geq km(p', u) + (1 - k)m(p'', u)$$

for given u .

Proof: Let x' and x'' solve the expenditure minimization problem when the price vector is respectively p' and p'' . By definition of the expenditure function, $p'x' = m(p', u)$ and $p''x'' = m(p'', u)$. Likewise, let \bar{x} solve the problem when the price vector is \bar{p} , so that $\bar{p}\bar{x} = m(\bar{p}, u)$. Since x' and x'' are solutions to their respective expenditure minimization problems we must have

$$p'\bar{x} \geq p'x' \quad \text{and} \quad p''\bar{x} \geq p''x'' \quad [\text{A.8}]$$

Multiplying through the first inequality by k and the second by $1 - k$ and summing, gives

$$kp'\bar{x} + (1-k)p''\bar{x} \geq kp'x' + (1-k)p''x'' \quad [\text{A.9}]$$

But by definition of \bar{p} this implies

$$(kp' + (1-k)p'')\bar{x} = \bar{p}\bar{x} \geq kp'x' + (1-k)p''x'' \quad [\text{A.10}]$$

which is the result we want.

Figure 4.2 illustrates the proof of this important result. It is obvious that when the isoexpenditure lines at which x' and x'' are respectively optimal solutions are shifted so as to pass through point \bar{x} , they must yield higher expenditure, thus giving the key inequalities in [A.8]. The rest of the proof then follows by simple algebra. The figure could in one sense be misleading. The inequalities (which in this case are *strict*) appear to follow from the convexity of the indifference curves. Note, however, that the inequalities in [A.8] follow simply from the fact that x' (resp. x'') minimizes px at price vector p' (resp. p'') while \bar{x} may not – [A.8] then follows from the definition of a minimum. Thus the proof of concavity of the expenditure function does *not* depend on convexity of preferences. However, the property of *uniqueness* of solutions like x' and x'' , and the differentiability of Hicksian demands and of the expenditure function, do. Note that strict convexity of preferences implies strict concavity of the expenditure function at an interior solution to problem [A.1], since it implies uniqueness of the solution and hence strict inequalities in [A.8].

Figure 4.3 illustrates the strict concavity of the expenditure function when the price vectors p' and p'' differ only in respect of one price, p_i . We can now show that the slope of the expenditure function at a point is equal to the compensated demand for good i at the price p_i :

(b) *Shephard's Lemma.* $\partial m(p, u)/\partial p_i = x_i^* = H_i(p, u)$. The proof is just a version of the Envelope Theorem of section 2J. First, differentiate through [A.7] to obtain

$$\frac{\partial m}{\partial p_i} = x_i^* + \sum_{j=1}^n p_j \frac{\partial x_j^*}{\partial p_i} \quad i = 1, \dots, n \quad [\text{A.11}]$$

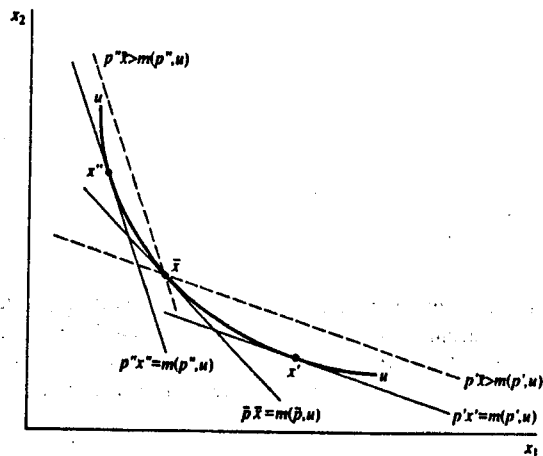


Fig. 4.2

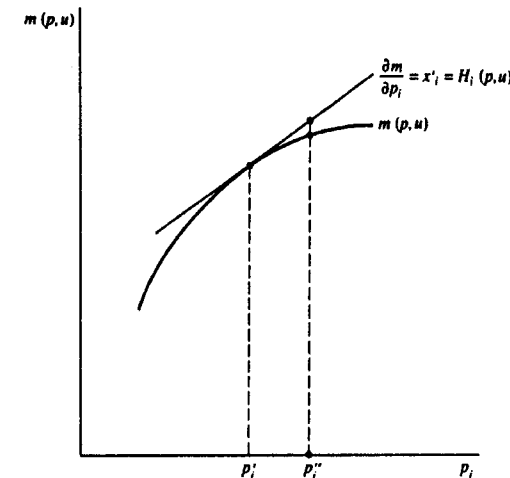


Fig. 4.3

From the first-order conditions [A.3] we have $p_j = \mu u_j$ and so we need to establish that $\mu \sum_j u_j \partial x_j^* / \partial p_i = 0$. But differentiating through the constraint [A.4] w.r.t. p_i , with u constant, does this.

Thus the partial derivative of the expenditure function with respect to the i th price is the compensated demand for the i th good. In Fig. 4.3, the slope of the curve at price p'_i is $x'_i = H_i(p_1, \dots, p'_i, \dots, p_n, u)$. This can be put intuitively as follows. Suppose a consumer buys 12.5 units of gas a week at a cost of £1 per unit. The price of gas then rises by 1p per unit. Shephard's lemma says that, *to a first approximation*, to maintain the same utility level or standard of living her expenditure must increase by 12.5p: just enough to maintain consumption at the initial price level. The qualification 'to a first approximation' is important. This is only strictly true in the limit, as the price change tends to zero. For *finite* price changes Fig. 4.3 shows that this overstates the required increase in expenditure, since the expenditure function is strictly concave. The reason for this is clear from Figs 4.2 and 4.3: as a good's price goes up, the consumer substitutes *away* from the good in question, and this reduces the amount of expenditure otherwise required to keep utility constant. Recall from Chapter 3 the distinction between Hicks and Slutsky compensated demands. Shephard's lemma tells us that for small enough price changes this distinction can be ignored.

(c) $\partial m / \partial p_i \geq 0$ with strict inequality if $x_i^* > 0$. This follows immediately from Shephard's lemma. Since at least one good must be bought, the expenditure function is non-decreasing in the price vector p and strictly increasing in at least one price. Higher prices mean higher expenditure to reach a given utility.

(d) *The expenditure function is homogeneous of degree 1 in prices.* This is easily seen. Take a given u value and price vector p^0 , and let $m^0 = m(p^0, u) = p^0 x^0$ where x^0 is the expenditure-minimizing bundle at p^0 . Then with a new price vector kp^0 , $k > 0$, relative

prices are unchanged and so, with u unchanged, x^0 must again solve the expenditure minimizing problem. Hence $m(kp^0, u) = kp^0 x^0 = km^0$.

(e) *The expenditure function is increasing in u .* Higher utility at given prices requires higher expenditure. Rather than use the envelope theorem again, recall that the Lagrange multiplier $\mu > 0$ in [A.2] is equal to the derivative $\partial m / \partial u$ (see section 2G). μ is the 'marginal cost of utility', since it represents the rate of change of *minimized* expenditure with respect to the required utility level. It is straightforward to confirm that μ is the reciprocal of the Lagrange multiplier λ in the corresponding utility maximization problem, i.e. μ is the inverse of the 'marginal utility of income' (see Question 3, Exercise 4A). Note, however, that although the assumptions underlying ordinal utility theory allow the sign of μ to be established, we cannot say that μ is necessarily increasing, or decreasing, with u , because both are possible for different, permissible utility functions (see Question 3, Exercise 4A).

It is important to be clear about the relation between expenditure and utility. Recall that the essential facts about the consumer's preference ordering are contained in the structure of her indifference sets or curves. The minimum expenditure required to reach a given indifference set at given prices is quite unaffected by any number we attach to that indifference set to indicate its place in the ordering. On the other hand, once we have chosen a numerical representation of the preference ordering – a utility function – this will imply a particular relationship between expenditure m and utility u . But the properties we set out above hold for all permissible utility functions, and the only general restriction we can place on the relation between m and u (for a given price vector) is that it is monotonically increasing.

Exercise 4A

1. *Cobb–Douglas utility function.* A consumer has the utility function $u = x_1^a x_2^b$, $a + b = 1$. Derive her Hicksian demand functions and expenditure function. Confirm that the expenditure function possesses the properties set out in this section. Then derive the expenditure function for the utility function $v = u^2$ and compare it to the one you obtained previously. In particular, compare the values $\partial m / \partial u$ in each case.
2. If goods are perfect complements the consumer's utility function can be written: $u = \min(x_1, x_2)$. If the goods are perfect substitutes the utility function can be written $u = ax_1 + bx_2$. Discuss the nature of the expenditure function in each of these cases.

3* Consider the problems:

$$\max u(x) \text{ s.t. } px = m; \quad \min px \text{ s.t. } u = u(x)$$

where u is a strictly quasi-concave utility function and p is the same price vector in each case. With m given, let u^* be the optimized utility in the first problem, with every $x_i^* > 0$. Then let u^* be the value of the utility constraint in the second problem. Then show:

- (a) the solution vector in the second problem is identical to that in the first;
- (b) $\lambda^* = 1/\mu^*$, where λ^* and μ^* are the optimal values of the Lagrange multipliers in the first and second problems respectively;

(c) these results hold for any positive monotonic transformation of the utility function.

4. *Stone–Geary utility function.* A consumer has the utility function $u = (x_1 - c_1)^a (x_2 - c_2)^b$, $a + b = 1$, where the c_i are interpreted as minimum subsistence levels of x_i , $i = 1, 2$. Derive the consumer's Hicksian demand functions and expenditure function, and compare them to the results obtained in Question 1.

B. The indirect utility function, Roy's identity and the Slutsky equation

The indirect utility function is derived from the consumer problem of maximizing $u(x_1, \dots, x_n)$ subject to the budget constraint $\sum p_i x_i \leq M$ and non-negativity constraints. We saw in section 3D that the x_i which are optimal for this problem will be functions of the p_i and M : $x_i^* = D_i(p_1, \dots, p_n, M) = D_i(p, M)$. The *maximized* value of $u(x_1, \dots, x_n) = u(x_1^*, \dots, x_n^*)$ will therefore also be a function of the p_i and M :

$$\begin{aligned} u(x_1^*, \dots, x_n^*) &= u(D_1(p, M), \dots, D_n(p, M)) \\ &= u^*(p, M) \end{aligned} \quad [\text{B.1}]$$

u^* is known as the *indirect utility function* since utility depends indirectly on prices and money income via the maximization process, in contrast to the normal utility function $u(x_1, \dots, x_n)$ where u depends directly on the x_i . We can use u^* to investigate the effects of changes in prices and money income on the consumer's utility.

From the interpretation of the Lagrange multiplier, the effect of an increase in money income on the maximized utility is

$$\frac{\partial u^*}{\partial M} = \lambda \quad [\text{B.2}]$$

The effect of a change in p_i on u^* can be found as a version of the Envelope Theorem. Differentiating u^* with respect to p_i :

$$\frac{\partial u^*}{\partial p_i} = \sum u_k \frac{\partial x_k^*}{\partial p_i} = \lambda \sum p_k \frac{\partial x_k^*}{\partial p_i} \quad [\text{B.3}]$$

The budget constraint must still be satisfied so that

$$\frac{d}{dp_i} \left(\sum p_k x_k^* \right) = \frac{dM}{dp_i} = 0$$

and so

$$\sum p_k \frac{\partial x_k^*}{\partial p_i} + x_i^* = 0$$

or

$$-x_i^* = \sum p_k \frac{\partial x_k^*}{\partial p_i}$$

Substitution of this in [B.3] gives Roy's identity:

$$\frac{\partial u^*}{\partial p_i} = -\lambda x_i^* = -\frac{\partial u^*}{\partial M} x_i^* \quad [\text{B.4}]$$

The expression on the right-hand side of [B.4] has the following intuitive explanation. An increase in p_i is a reduction in the purchasing power of the consumer's money income M , and by Shephard's lemma, to the first order, her purchasing power falls at the rate $-x_i^*$ as p_i varies. λ is the marginal utility of money income. The product of λ and $-x_i^*$ is the rate at which utility varies with money income, times the rate at which (the purchasing power of) money income varies with p_i , and so this product yields the rate of change of utility with respect to p_i .

Since $\lambda > 0$, Roy's identity shows, as we would expect, that an increase in the price of a good a consumer buys reduces her (maximized) utility or standard of living.

The indirect utility function tells us that utility depends, via the maximization process, on the price-income situation the consumer faces. Note that [B.2] implies that the indirect utility function is monotonically increasing in income, M . Thus we can invert the indirect utility function $u = u^*(p, M)$ to obtain the expenditure function $M = m(p, u)$. A given solution point for a given price vector can be viewed equivalently as resulting from minimizing expenditure subject to the given utility level or maximizing utility subject to the given expenditure level. We can choose either to solve the utility maximization problem, obtain the indirect utility function and invert it to obtain the expenditure function, or to obtain the expenditure function and then invert to obtain the indirect utility function (see Question 3, Exercise 4B). The two functions are *dual to each other*, and contain essentially the same information: the forms of the functions and their parameters are completely determined by the form of the original (direct) utility function. But then, since each of these three functions contains the same information, we can choose any one of them as the representation of the consumer's preferences that we wish to work with.

This duality can be used to give a neater derivation of Roy's identity. Setting $M = m(p, u)$, re-write the indirect utility function as

$$u = u^*(p, m(p, u)) \quad [\text{B.5}]$$

Then differentiating through w.r.t. p_i , allowing m to vary in such a way as to hold u constant, gives

$$0 = \frac{\partial u^*}{\partial p_i} + \frac{\partial u^*}{\partial M} \frac{\partial m}{\partial p_i} \quad [\text{B.6}]$$

which, using Shephard's lemma, and [B.2] gives Roy's identity [B.4] directly.

The Slutsky equation

The Slutsky equation plays a central role in analysing the properties of demand functions. It is derived as follows. If we take as the constraint in the utility maximization problem the level of expenditure resulting from solution of the expenditure minimization problem (or equivalently take as the constraint in the latter problem the level of utility resulting from the solution to the former) then the solutions x_i^* to the two problems, the values of

the Marshallian and Hicksian demand functions, will be identical. We can then write for the i th good

$$H_i(p, u) = D_i(p, M) = D_i(p, m(p, u)) \quad [\text{B.7}]$$

Since [B.7] is an identity we can differentiate through with respect to the j th price, allowing expenditure to change in whatever way is required to keep utility constant, to obtain

$$\frac{\partial H_i}{\partial p_j} = \frac{\partial D_i}{\partial p_j} + \frac{\partial D_i}{\partial M} \frac{\partial m}{\partial p_j} \quad i, j = 1, \dots, n \quad [\text{B.8}]$$

Using Shephard's lemma and rearranging gives the Slutsky equation

$$\frac{\partial D_i}{\partial p_j} = \frac{\partial H_i}{\partial p_j} - x_j \frac{\partial D_i}{\partial M} \quad i, j = 1, \dots, n \quad [\text{B.9}]$$

Taking $i = j$, so that we consider the effect of a price change on its own demand, we see from [B.9] that the slope of the Marshallian demand function is the sum of two effects: the *substitution effect*, $\partial H_i / \partial p_i$, which is the slope of the Hicksian or compensated demand curve; and the *income effect*, $-x_i \partial D_i / \partial M$. Thus the Slutsky equation gives a precise statement of the conclusions of the diagrammatic analysis of Chapter 3. We show in a moment that $\partial H_i / \partial p_i \leq 0$. Then [B.9], again with $i = j$, establishes that if the good is normal, so that $\partial D_i / \partial M > 0$, the slope of its Marshallian demand curve is negative. If the good is inferior, so that $\partial D_i / \partial M \leq 0$, the slope is negative, positive or zero depending on the relative sizes of the absolute values $|\partial H_i / \partial p_i|$ and $|x_i \partial D_i / \partial M|$.

It is useful to express the Slutsky equation in elasticity form. Again taking $i = j$, multiplying through [B.9] by p_i / x_i , and the income term by M / M , gives

$$\varepsilon_{ii} = \sigma_{ii} - s_i \eta_i \quad i, j = 1, \dots, n \quad [\text{B.10}]$$

where ε_{ii} is the Marshallian demand elasticity, σ_{ii} is the Hicksian or compensated demand elasticity, η_i is the income elasticity of demand, and $s_i = p_i x_i / M$ is the share of good i in total expenditure. Thus the difference between Hicksian and Marshallian elasticities for a good will be smaller, the smaller its income elasticity and the less significant it is in the consumer's budget. With $i \neq j$, [B.9] becomes

$$\varepsilon_{ij} = \sigma_{ij} - s_j \eta_i \quad [\text{B.11}]$$

which emphasizes that cross-price Marshallian demand elasticities depend both on compensated elasticities and on income elasticities weighted by expenditure shares. Equality of the Marshallian cross-price elasticities therefore requires strong restrictions on preferences (see Question 5, Exercise 4B).

We define the Slutsky matrix as the $n \times n$ matrix $[\partial H_i / \partial p_j]$ of Hicksian demand derivatives. It is a straightforward extension of Shephard's lemma and the properties of the expenditure function to show that this matrix is a *symmetric, negative semi-definite* matrix. Thus, since from Shephard's lemma

$$\frac{\partial m(p, u)}{\partial p_i} = H_i(p, u) \quad i = 1, \dots, n$$

we have

$$\frac{\partial^2 m(p, u)}{\partial p_j \partial p_i} = \frac{\partial H_i}{\partial p_j} \quad i, j = 1, \dots, n \quad [\text{B.12}]$$

Then, from *Young's Theorem*¹ we have immediately that $\partial H_i / \partial p_j = \partial H_j / \partial p_i$, and so the Slutsky matrix is symmetric. The Slutsky matrix $[\partial H_i / \partial p_j]$ is the matrix of second-order partials of the expenditure function and the concavity of the expenditure function implies that matrix is negative semi-definite. Since $\partial^2 m / \partial p_i^2 = \partial H_i / \partial p_i \leq 0$, by the definition of negative semi-definiteness (see section 2I), the Hicksian demand curve cannot have a positive slope. We have seen earlier that strict convexity of preference and $x_i > 0$ at the optimum, establish the stronger result that $\partial^2 m / \partial p_i^2 = \partial H_i / \partial p_i < 0$.

The Hicksian demand derivative $\partial H_i / \partial p_j$ is often used to define complements and substitutes. Two goods i and j are called *Hicksian complements* if $\partial H_i / \partial p_j < 0$ and *Hicksian substitutes* if $\partial H_i / \partial p_j > 0$. The advantage of this definition is that symmetry implies that the nature of the complementarity or substitutability between the goods cannot change if we take $\partial H_j / \partial p_i$ rather than $\partial H_i / \partial p_j$. This would *not* be true if we defined complements and substitutes in terms of the Marshallian demand derivatives (see Question 5, Exercise 4B).

Properties of demand functions

We have seen that it is possible to draw definite conclusions about the effects of price changes on the Hicksian demands. The Hicksian demand functions are not, however, directly observable since they depend on the consumer's utility level as well as prices. On the other hand, the Marshallian demand functions can be estimated from information on purchases, prices and money income but the theory yields few definite predictions about the effects of changes in prices and money income. The Slutsky equation enables us to reformulate the predictions about the properties of Hicksian demand functions in terms of the observable Marshallian demand functions and thus to widen the set of testable predictions from consumer theory.

We can summarize the testable implications derived in this and the previous chapter:

- Marshallian demand functions are homogeneous of degree zero in prices and money income;
- the Marshallian demand functions satisfy the 'adding up' property: $\sum p_i x_i^* = M$;
- the Hicksian demand derivatives (cross-substitution effects) are symmetric: $\partial H_i / \partial p_j = \partial H_j / \partial p_i$ or, using the Slutsky equation, $\partial D_i / \partial p_j + x_j \partial D_i / \partial M = \partial D_j / \partial p_i + x_i \partial D_j / \partial M$;
- the Slutsky matrix $[\partial H_i / \partial p_j] = [\partial D_i / \partial p_j + x_j \partial D_i / \partial M]$ is negative semi-definite.

These are all the predictions about the Marshallian demand functions which can be made on the basis of the consumer preference axioms. (As we will see in sections D and E, more detailed predictions require stronger and less general specifications of preferences.) The converse question of whether a system of demand functions with these properties implies the existence of a utility function from which the demand functions could have been

derived is known as the *integrability problem*. In the next section we will show that this is in fact so by considering the equivalent problem of retrieving an expenditure function (which also can be used to represent preferences) from a set of Marshallian demand functions which satisfy the above properties.

Exercise 4B

- Show that the Hicksian demand function is homogeneous of degree zero in prices. Then, use the fact (Euler's Theorem) that if a function $f(x_1, \dots, x_n)$ is homogeneous of degree zero, we have $\sum_{i=1}^n f_i x_i = 0$, to prove that $\sum_{j=1}^n (\partial H_i / \partial p_j) p_j = 0$. Interpret this in terms of the Slutsky matrix.
- The consumer has the utility function $u = x_1^\alpha x_2^{1-\alpha}$. Find her indirect utility function. Confirm Roy's identity by:
 - differentiating the indirect utility function with respect to the price of good 1;
 - using the first-order conditions to obtain solutions for x_1 and λ , and therefore an expression for $-\lambda x_1$;
 - showing that (a) and (b) give the same result.
- Invert the indirect utility function you obtain in Question 2 to express expenditure as a function of prices and utility. Then show that this is the expenditure function for this form of direct utility function.
- (a) Show that the Marshallian demand functions satisfy the following restrictions:

$$\text{Cournot aggregation: } \sum_{i=1}^n p_i \frac{\partial D_i}{\partial p_j} + x_j = 0 \quad j = 1, \dots, n$$

$$\text{Engel aggregation: } \sum_{i=1}^n p_i \frac{\partial D_i}{\partial M} = 1$$

(Hint: use the adding up property and differentiate.)

- Express these restrictions in elasticity form

$$\sum_i s_i e_{ij} + s_j = 0 \quad j = 1, \dots, n$$

$$\sum_i s_i \eta_i = 1$$

where e_{ij} is the cross-price elasticity $(\partial D_i / \partial p_j)(p_j / x_i)$, η_i is the income elasticity $(\partial D_i / \partial M)(M / x_i)$ and $s_i = p_i x_i / M$ is the budget or expenditure share of the i th good.

- Show that the homogeneity property implies

$$\sum_{i=1}^n p_i \frac{\partial D_i}{\partial p_j} + M \frac{\partial D_i}{\partial M} = 0 \quad i = 1, \dots, n$$

and express this in elasticity form

$$\sum_i e_{ij} + \eta_i = 0 \quad i = 1, \dots, n$$

- Show that if a set of Marshallian demand functions satisfies homogeneity, symmetry, and Engel aggregation they will also satisfy Cournot aggregation.

5. Show that if complements and substitutes are defined in terms of Marshallian demand derivatives, goods could be, say, complements on the basis of the sign of $\partial D_i / \partial p_j$, and substitutes on the basis of the sign of $\partial D_j / \partial p_i$. Give precise conditions under which this occurs.
6. Show that if the utility function $u(x)$ is an ordinal representation of preferences no restrictions can be placed on the signs of $\partial^2 u^*(p, M) / \partial M^2 = \partial \lambda / \partial M$ and $\partial^2 u^*(p, M) / \partial M \partial p_i = \partial \lambda / \partial p_i$. (Hint: consider positive monotonic transformations of the utility function $G(x) = g(u(x))$, $g' > 0$) Interpret the result. Is it possible to find a numerical representation of preferences $u(x)$ such that the marginal utility of income $\partial u^* / \partial M$ is constant with respect to all prices and income?
7. *Quasi-convexity in prices of the indirect utility function.* Sketch price indifference curves in p_1, p_2 space which show combinations of prices which yield the consumer the same maximized utility level: i.e. sketch the contours of $u^*(p_1, p_2, M)$. What is the slope of the indifference curve? (Hint: use Roy's identity.) Show that it is unaffected by positive monotonic transformations of the direct utility function. Show that $u^*(p, M)$ is quasi-convex in prices: the set of p such that $u^*(p, M) \leq u^0$ is convex. Then show that if two price vectors p^0, p^1 yield the same utility $u^*(p^0, M) = u^*(p^1, M)$, the consumer would prefer the risky prospect of facing (p^0, M) or (p^1, M) with equal probabilities to facing the certain prospect (\bar{p}, M) where $\bar{p} = \frac{1}{2}p^0 + \frac{1}{2}p^1$. Does this imply that consumers are made worse off by price stabilization?

C. Measuring the benefits of price changes

We often wish to measure the benefit to consumers of a change in the price of a commodity. The price change may result, for example, from changes in tariffs on imported goods, or in the rate of purchase tax and we may want to estimate the effects of these on consumers' welfare for public policy purposes. We know that a change in a price will alter the feasible set confronting a consumer, that a new optimal bundle of goods will result, and that the consumer will be on a new indifference curve. In the case of a price fall the consumer will be better off in the sense that he prefers the new bundle to the initial one. How can we measure this benefit? One suggestion might be by the change in the utility level of the consumer. This suffers from a number of serious drawbacks, chief among which is the fact that no significance attaches to the size of utility differences, only to their sign. This means that a utility measure would be essentially arbitrary. Furthermore, any utility measure would not be comparable among different individuals and we could not add utility differences for a measure of total benefit to all consumers.

A measure which at least avoids this last problem is the consumer's own monetary valuation of the price change. Since the measure is expressed in terms of money, individual measures are at least commensurable and could in principle be added to form a measure of the aggregate benefit to all consumers of the good.

We stress 'in principle' because if the aggregate monetary measure is to be used for policy purposes, an important value judgement must be made before the individual monetary measures can be summed. This is that an extra £1 of benefit to an individual has the same social significance to whichever individual it accrues. This becomes particularly

important in cost-benefit analysis when some individuals gain and others lose as a result of particular decisions. Then we have to make the value judgement that £1 of benefit to one individual can offset £1 of loss to another.

Figure 4.4 illustrates the effect of a fall in the price of good 1 from p_1^0 to p_1^1 with money income and the price of good 2 held constant. The consumer's initial bundle is A on I_0 and the bundle chosen after the fall in p_1 is B on I_1 . The consumer is better off, but what is his monetary valuation of this change in utility? One answer is the maximum amount he would be prepared to pay for the opportunity of buying good 1 at the new price rather than at the old price. This is the *compensating variation (CV)* measure and is defined formally as the amount of money which must be taken from the consumer in the new situation in order to make him as well off as he was in the initial situation. It is identical to the compensating variation in money income used in section 3D to decompose the price effect into income and substitution effects. Notice that the definition used here applies equally well to price rises, in which case the compensating variation takes the opposite sign: the consumer becomes worse off and must be given money to make him as well off with the new prices as he was with the old.

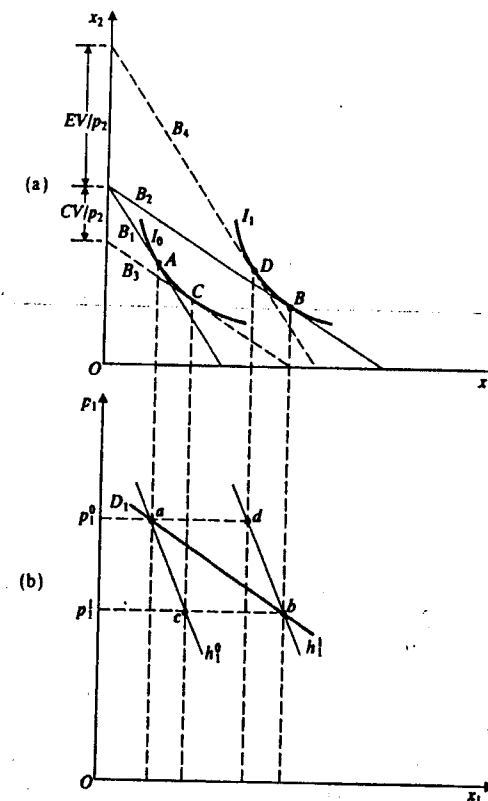


Fig. 4.4

The *CV* measure is not the only plausible monetary measure of the gain to the consumer of a change in the price of a good. The *equivalent variation (EV)* is the amount of money which would have to be given to the consumer when he faces the initial price, to make him as well off as he would be facing the new price with his initial income. Again the definition allows for a rise as well as a fall in price. Both the *CV* and *EV* definitions allow for more than one price to change at the same time, but we will restrict ourselves for the moment to the case of a single price change.

The *EV* and *CV* are shown in Fig. 4.4(a). *CV* is the change in M required to shift the budget line from B_2 to B_3 so that the consumer's utility level after the price fall is the same as it was before. *CV* is equal to p_2 times the difference in the x_2 intercept of B_2 and B_3 . *EV* is the change in M required to shift the budget line from B_1 to B_4 so that facing the initial prices he can just achieve the utility level he would have had with the new prices. *EV* is equal to p_2 times the difference in the x_2 intercept of B_1 and B_4 . Notice that *CV* is not in general equal to *EV*.

The distinction between *EV* and *CV* can be expressed in terms of the indirect utility function introduced in section B. In the initial situation the consumer faces prices $p^0 = (p_1^0, \dots, p_n^0)$ with income M^0 and his maximized utility is $u^*(p^0, M^0) = u^0$. With the new prices $p^1 = (p_1^1, \dots, p_n^1)$ and the same income his maximized utility becomes $u^*(p^1, M^0) = u^1$. *CV* is the change in money income necessary to make his utility when he faces p^1 equal to the initial utility level when he faced p^0 with an income of M^0 . Hence *CV* is defined by

$$u^*(p^0, M^0) = u^*(p^1, M^0 - CV) = u^0 \quad [\text{C.1}]$$

EV is the change in M necessary to make his utility when facing p^0 equal to his utility when facing p^1 with income of M^0 . *EV* is therefore defined by

$$u^*(p^0, M^0 + EV) = u^*(p^1, M^0) = u^1 \quad [\text{C.2}]$$

We can also define *CV* and *EV* by using the expenditure function introduced in section A. The minimum level of expenditure necessary to achieve the consumer's initial utility level u^0 with the initial price vector p^0 is $m(p^0, u^0) = M^0$. The minimum level necessary to achieve this initial utility level when prices alter to p^1 is $m(p^1, u^0)$, so that the difference between $m(p^0, u^0)$ and $m(p^1, u^0)$ is the change in income necessary to ensure that the consumer is indifferent between facing prices p^0 with income M^0 and prices p^1 with a different income. This change in income however is just the compensating variation so that:

$$CV = M^0 - m(p^1, u^0) = m(p^0, u^0) - m(p^1, u^0) \quad [\text{C.3}]$$

If only one price, say p_1 , falls from p_1^0 to p_1^1 we must have

$$m(p^0, u^0) - m(p^1, u^0) = \int_{p_1^1}^{p_1^0} \frac{\partial m}{\partial p_1} dp_1$$

But we saw in section A that $\partial m / \partial p_1 = x_1^* = H_1(p, u^0)$ and so

$$CV = m(p^0, u^0) - m(p^1, u^0) = \int_{p_1^1}^{p_1^0} H_1(p, u^0) dp_1 \quad [\text{C.4}]$$

$H_1(p, u^0)$ is the Hicksian constant utility demand function for x_1 , and if all other prices are held constant we can draw, as in Fig. 4.4(b), the constant utility demand curve h_1^0 , showing the relationship between p_1 and x_1 when utility is constant at $u = u^0$. For a price fall *CV* is the area $p_1^0 a c p_1^1$.

The consumer's market demand curve for x_1 is not, however, his constant utility demand curve but rather his constant money income demand curve, D_1 . But from the Slutsky equation we saw that since the constant utility demand curve plots the substitution effect of a price change and the constant money income demand curve plots the whole price effect (i.e. the substitution and income effects) the two curves will coincide if and only if the income effect is zero. Equivalently, the consumer's indifference curves in Fig. 4.4(a) must be vertically parallel.

When D_1 and h_1^0 coincide *CV* is the area between the price lines p_1^0 and p_1^1 under the consumer's market demand curve. If the income effect is non-zero then the area under the consumer's market demand curve between the price lines will not be equal to *CV*. In particular if x_1 is a normal good ($\partial D_1 / \partial M > 0$) then D_1 will exceed h_1^0 for all $p_1 < p_1^0$ and the area under the D_1 curve between the price lines will exceed *CV*, as Fig. 4.4(b) illustrates.

Points A, B, C, in Fig. 4.4(a) correspond to points a, b, c in Fig. 4.4(b) and D_1 cuts h_1^0 at a. If x_1 had been an inferior good then D_1 would have been below h_1^0 for $p_1 < p_1^0$ and *CV* would have been underestimated by the area under the D_1 curve between the price lines.

A similar approach can be used for *EV*. The value of the expenditure function $M^0 = m(p^1, u^1)$ is the minimum expenditure necessary to achieve the new post-price change utility level and $m(p^0, u^1)$ is that necessary to achieve the new level of utility with the initial prices. Hence in the case of a price fall from p_1^0 to p_1^1 :

$$EV = m(p^0, u^1) - m(p^1, u^1) = \int_{p_1^1}^{p_1^0} H_1(p, u^1) dp_1 \quad [\text{C.5}]$$

In Fig. 4.4(b) h_1^1 is the constant utility demand curve for $u = u^1$ and *EV* is the area under h_1^1 and between the price lines p_1^0, p_1^1 . Since the income effect is non-zero, h_1^1 and D_1 intersect at b and the area under D_1 between the price lines is an underestimate of *EV*.

We can relate this discussion to the idea of *consumer surplus*. In early attempts to associate measures of consumer welfare with areas under demand curves, it was argued by the French engineer J. Dupuit and the English economist A. Marshall that the area under an individual's constant money income (Marshallian) demand curve up to the quantity being consumed gave a money measure of the benefit of that consumption. Subtracting the expenditure on the good from this area then gave the net benefit, or consumer surplus, derived from the good. We can examine this idea in the light of the consumer theory of this chapter.

Consider the consumer's indirect utility function $u^*(p_1, \dots, p_n, M)$, and let p_1^0 now denote the lowest price at which, given the remaining prices p_2, \dots, p_n , the consumer's demand for good 1 is just zero. The actual price of good 1 is denoted p_1^1 . Roy's identity gives

$$\frac{\partial u^*}{\partial p_1} = -\lambda x_1 = -\lambda D_1(p_1, \dots, p_n, M)$$

where λ is the marginal utility of income. Integrating over the interval $[p_1^0, p_1^1]$ gives

$$\int_{p_1^0}^{p_1^1} \frac{\partial u^*}{\partial p_1} dp_1 = -\lambda \int_{p_1^0}^{p_1^1} D_1(p_1, \dots, p_n, M) dp_1 \quad [\text{C.6}]$$

if and only if λ can be treated as a constant when p_1 changes. Thus we have

$$\frac{1}{\lambda} [u^*(p_1^1, \dots, p_n, M) - u^*(p_1^0, \dots, p_n, M)] = \int_{p_1^0}^{p_1^1} D_1(p_1, \dots, p_n, M) dp_1 \quad [\text{C.7}]$$

The left-hand side can be regarded as a *money measure* of the change in utility caused by a change in price from p_1^0 to p_1^1 (since λ is in units of utility/£ while u^* is in units of utility), while the right-hand side is the area under the Marshallian demand curve for good 1 between the prices p_1^0 and p_1^1 .

Unfortunately it is in general not the case that a consumer's preferences can be represented by a utility function $u(x)$ such that the marginal utility of money income $\partial u^*/\partial M = \lambda$ is constant when a price changes. Using Roy's identity, we see that

$$\frac{\partial \lambda}{\partial p_i} = \frac{\partial^2 u^*}{\partial M \partial p_i} = \frac{\partial(-\lambda x_i)}{\partial M} = -x_i \frac{\partial \lambda}{\partial M} - \lambda \frac{\partial x_i}{\partial M} = 0 \quad [\text{C.8}]$$

is necessary and sufficient for λ to be constant with respect to p_i . Multiplying through [C.8] by $M/\lambda x_i$ we can express the condition as

$$\eta_i = -\rho \quad [\text{C.9}]$$

where η_i is the income elasticity of demand for good i and ρ is the elasticity of marginal utility of income. It is possible to specify preferences which can be represented by utility functions which satisfy [C.8] or [C.9]. In fact in the example in Question 2 we have $\partial x_i/\partial M = 0 = \partial \lambda/\partial M$. However, [C.8] or [C.9] greatly restrict the preference orderings for which it is valid to use the area under the Marshallian demand curve $D_i(p, M)$ from p_i^0 to p_i^1 as a money measure of the change in utility arising from a change in p_i^1 from p_i^0 to p_i^1 . The difficulties with using the areas under the Marshallian demand curves as welfare measures are compounded if more than one price changes. (See question 4.)

We do, however, have money measures of benefits which do *not* require such restrictive assumptions, namely the *CV* and *EV*. In Fig. 4.5(b), h_1^0 and h_1^1 are the Hicksian demand curves corresponding to the pre- and post-price change utilities in Fig. 4.5(a), and D_1 the corresponding Marshallian demand. In exactly the same way as before, we can show that *CV* is given by the area $p_1^0 c p_1^1$ and *EV* by the area $p_1^0 d b p_1^1$. All that differs is that in the initial equilibrium, $x_1 = 0$.

There would seem to be two problems with the Hicksian measures of consumer surplus. One is that they are not unique – *CV* and *EV* in general differ, so which is 'right'? The other is that the Hicksian demand functions are not directly observable from market data, so how are *CV* and *EV* to be made operational?

The difference between *CV* and *EV* is inescapable without severe restrictions on preferences. If the income effect is not zero then the answer to the question: how much income can we take from the consumer to cancel out the welfare gain resulting from the fall in price from p_1^0 to p_1^1 ? is bound to differ from the answer to the question: how much

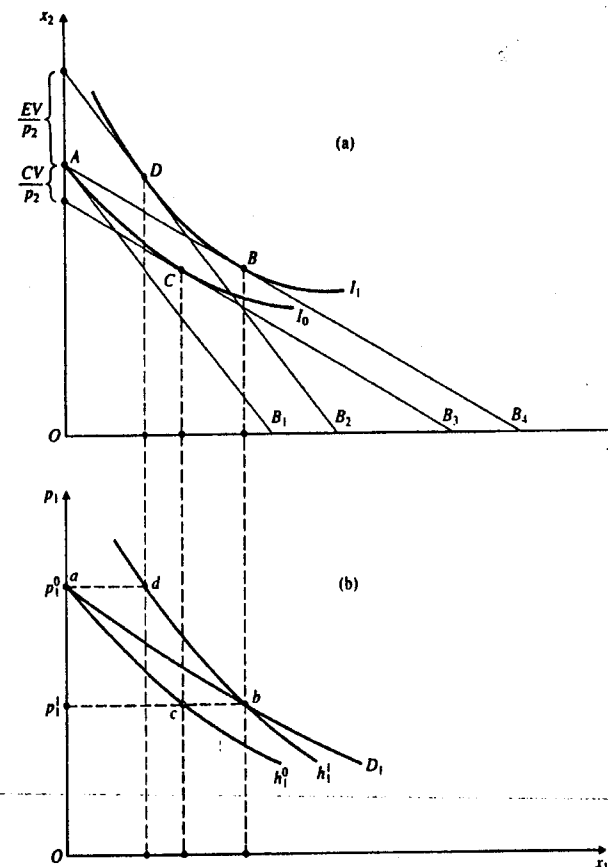


Fig. 4.5

income must we give the consumer to make him just as well off with price p_1^0 as he would be with price p_1^1 ?

Which measure, *CV* or *EV*, is appropriate, depends on which is the relevant question when a money measure of benefit is required. If, for example, a public project having the effect of reducing price of good x_1 from p_1^0 to p_1^1 is to be financed by a lump-sum tax on each consumer, we can say that the project should certainly go ahead if each consumer's *CV* exceeds the tax she has to pay, since in that case she must be better off as a result of the project. Another example is when a government might wish to consider whether to pay a subsidy to producers to reduce the price of good 1 from p_1^0 to p_1^1 . If the cost of the subsidy per consumer (including administrative costs) exceeds each consumer's corresponding *EV* then the government would do better by making a lump-sum payment to each consumer equal to her *EV*, since the same utility gain is achieved at a lower total cost.

In the discussion so far we have considered only changes in a single price and we might be interested in measuring the benefits of changes in any number of prices. The definitions of CV and EV in terms of the differences in values of the expenditure function apply equally to this case: in [C.3] and [C.5] we can let the price vectors p^0 and p^1 differ with respect to as many prices as we wish. In each case we are finding an income change which makes the consumer indifferent between the two price vectors, with the CV corresponding to the pre-change utility level and the EV to the post-change utility level.

The choice of measure depends on the purpose of the measurement, but how are we actually to measure CV s or EV s in any given context? One approach might be to argue that, since the Hicksian demands are not directly observable, we should take the relevant area under the consumer's Marshallian demand function as an approximation to the appropriate measure. If income effects for the good are very small, one can claim that the approximation will be close. However, we can show that if we have estimates of an individual's Marshallian demand functions, then such an approximation is unnecessary. Provided these functions satisfy the restrictions implied by consumer theory, the expenditure function can be 'retrieved' from the Marshallian demand functions, and once we have the expenditure function the CV and EV measures follow directly.

This can be proved by considering a version of a problem with a long history of study in economics, the *Integrability Problem*, the general form of which is as follows. Suppose that we have a given system of n partial differential equations

$$\frac{\partial y}{\partial x_i} = g_i(y, x) \quad i = 1, \dots, n \quad [\text{C.10}]$$

where the g_i are given functions, y is a real variable and x is a vector of n real variables. A solution to the system is a function $y = f(x)$ which satisfies the n equations as an identity, i.e. they hold for all values of x . Such a function exists if the *Hurwicz-Uzawa integrability condition*

$$g_j(y, x) \frac{\partial g_i}{\partial y} + \frac{\partial g_i}{\partial x_j} = g_i(y, x) \frac{\partial g_j}{\partial y} + \frac{\partial g_j}{\partial x_i} \quad i, j = 1, \dots, n \quad [\text{C.11}]$$

is satisfied for every pair of variables x_i, x_j .

We can apply [C.11] to our problem as follows. Suppose that we have estimated a system of Marshallian demand functions for the consumer

$$x_i = D_i(p, M) \quad i = 1, \dots, n \quad [\text{C.12}]$$

Taking the value of u as a fixed parameter, we can write

$$D_i(p, m(p, u)) = H_i(p, u) = \frac{\partial m}{\partial p_i} \quad i = 1, \dots, n \quad [\text{C.13}]$$

where we use Shephard's lemma. The expenditure function $m(p, u)$ is unknown, but the problem of finding it is precisely that of solving a system of the type [C.10], with the price vector p identified as the vector x , M as the variable y and m as the function f . Applying the integrability condition [C.11] we can solve [C.13] for the expenditure function if

$$x_j \frac{\partial D_i}{\partial M} + \frac{\partial D_i}{\partial p_j} = x_i \frac{\partial D_j}{\partial M} + \frac{\partial D_j}{\partial p_i} \quad i, j = 1, \dots, n \quad [\text{C.14}]$$

for all pairs of prices p_i, p_j (where we have used the fact that $x_i = D_i(p, M)$). But from the Slutsky equations, we see that [C.14] is precisely the condition that $\partial H_i / \partial p_j = \partial H_j / \partial p_i$, that is, that the Slutsky matrix be symmetric. Since this symmetry is implied by the theory, we conclude that we can obtain the consumer's expenditure function from the estimated Marshallian demand functions provided these also satisfy the restrictions implied by consumer theory: they must be homogeneous of degree zero in prices and income, satisfy the Slutsky equation, and satisfy the adding up condition that $\sum_i p_i D_i(p, M) = M$, so that expenditure just exhausts income at any price vector.

It may be no easy matter in practice actually to solve the given system of Marshallian demands for the expenditure function. In empirical demand analysis a simpler route is chosen. A particular functional form for the expenditure function (or equivalently the indirect utility function) is assumed, and the Marshallian demand functions corresponding to that form are estimated. It is then straightforward to retrieve the expenditure function parameters from the estimated equations. The main drawback is that the estimated functions may not be those that best fit the data in the standard statistical sense.

There is one important caveat to the conclusion that exact measures of the benefit of price changes, CV and EV , can be derived from knowledge of a consumer's Marshallian demand functions, so that approximations by areas under the Marshallian demand function are unnecessary. In many cases where we wish to evaluate the benefits of policies, we are only able to estimate the *aggregate market* Marshallian demand function, rather than individual consumer demand functions. It is not in general possible to use this to infer anything about the parameters of individual Hicksian demand functions. The conditions under which this can be done are very restrictive, and are discussed further in section E, in the context of the problem of aggregation.

Exercise 4C

1. Restate the analysis of Fig. 4.4 and the interpretation of CV and EV for the case in which the price of good 1 increases from p_1^0 to p_1^1 .

2. *Quasi-linear utility.* Suppose that the consumer's preferences can be represented by the quasi-linear utility function $u = f(x_1) + x_2$, $f' > 0$, $f'' < 0$.

(a) Show that the consumer's indifference curves are vertically parallel, i.e. their slope depends only on x_1 and not on x_2 .

(b) Set up and solve the consumer's utility maximization problem.

(c) Confirm that the income elasticity of demand for good 1 is zero and thus that the CV and EV for changes in p_1 are equal.

(d) Show that the marginal utility of income is independent of p_1 so that the change in Marshallian consumer surplus is a measure of the change in utility caused by changes in p_1 .

(e) What is the relationship between the change in the Marshallian consumer surplus and the EV and CV measures in this case?

3. Calculate the CV , EV and the change in the Marshallian consumer surplus for a consumer with preferences represented by the utility function $u = x_1 x_2$ with income

$M = 100$, $p_2 = 1$ and p_1 falling from 1 to $\frac{1}{2}$. Do this consumer's preferences satisfy the condition for the change in the Marshallian consumer surplus to be a valid measure of the change in utility?

4.* Multiple price changes.

- (a) When more than one price changes the Marshallian measure is not well defined in general because it depends on the order in which the prices are assumed to change: it is a path dependent line integral. It is path independent only if the cross-price demand effects are equal: $\partial D_i(p, M)/\partial p_j = \partial D_j(p, M)/\partial p_i$. Show that if the Marshallian measure is to be well defined for all possible price changes the consumer's preferences must be such that all income elasticities are unity. (Note that this implies that preferences are homothetic – see section D.) (Hint: use the Slutsky equation.)
- (b) Show that the CV and EV measures are well defined for all possible price changes without any restriction on consumer preferences.

D. Composite commodities, separability and homotheticity*

The analysis so far has developed the implications of the general set of assumptions on preferences and the budget constraint given in Chapter 3. We were able to place a number of restrictions on the forms of the demand and expenditure functions. However, for some purposes, especially applications of demand theory and empirical estimation of demand functions, further restrictions are useful. In this section we consider first an assumption about prices, and then some assumptions about the form of the utility or expenditure functions, which are useful in many circumstances.

Composite commodities

Suppose for example that we wish to analyse an individual's choice of labour supply and consumption goods. Although we could model her choice of the entire vector of consumption goods we are primarily interested in the trade-off between labour supply and 'consumption' in general. The only price in whose variations we are interested is the wage rate. It is then useful to treat the entire bundle of consumption goods as a single 'composite commodity'. The composite commodity theorem, due to J. Hicks, tells us that we can do this as long as we assume that the *relative prices* of the consumption goods remain constant throughout the analysis.

The composite commodity theorem. If the relative prices of a group of commodities x_1, x_2, \dots, x_g , $g \leq n$, are fixed, then they can be treated for purposes of demand analysis as a single composite commodity with a price given by an appropriate index of the prices of the goods p_1, \dots, p_g .

If the prices of the group of goods always move in proportion to each other then

$$p_2 = k_2 p_1, p_3 = k_3 p_1, \dots, p_g = k_g p_1 \quad \text{and} \quad k_i > 0, i = 2, \dots, g \quad [\text{D.1}]$$

for some constants k_i . Here the choice of good 1 as the 'group numeraire' is quite arbitrary. We can define the composite commodity as $x_c \equiv x_1 + \sum_{i=2}^g k_i x_i$ and we take as its price 'index' $p_c = p_1$ (see also Question 1, Exercise 4D). The idea of the theorem is that if we were to construct the consumer's preference ordering over consumption bundles consisting of the composite commodity and all other commodities, represent it by the utility function $u(x_c, x_{g+1}, \dots, x_n)$, and maximize this subject to the budget constraint $p_c x_c + \sum_{j=g+1}^n p_j x_j = M$, then we would obtain demand functions $D_c(p_c, p_{g+1}, \dots, p_n, M)$, $D_j(p_c, p_{g+1}, \dots, p_n, M)$; $j = g+1, \dots, n$, such that the D_j functions would be exactly those obtained from the corresponding problem with the original consumption bundle $(x_1, x_2, \dots, x_g, \dots, x_n)$, and the demand function for the composite commodity would be

$$D_c = D_1(p_1, \dots, p_g, \dots, p_n, M) + \sum_{i=2}^g k_i D_i(p_1, \dots, p_g, \dots, p_n, M)$$

Recall that working with the direct utility function, the indirect utility function or the expenditure function are equivalent ways of analysing consumer demands because they contain the same information about preferences. We can prove the composite commodity theorem by using the indirect utility function (for an approach based on the expenditure function see Question 1). Taking the n commodities individually the indirect utility function is $u = u^*(p_1, p_2, \dots, p_n, M)$. Using [D.1] we can write this as

$$\begin{aligned} u &= u^*(p_1, p_2, \dots, p_n, M) = u^*(p_1, k_2 p_1, \dots, k_g p_1, p_{g+1}, \dots, p_n, M) \\ &= v(p_1, p_{g+1}, \dots, p_n, M) = v(p_c, p_{g+1}, \dots, p_n, M) \end{aligned} \quad [\text{D.2}]$$

Hence applying Roy's identity we have

$$\begin{aligned} \frac{\partial v}{\partial p_c} &= \frac{\partial u^*}{\partial p_1} + \sum_{i=2}^g k_i \frac{\partial u^*}{\partial p_i} = -\lambda \left[x_1 + \sum_{i=2}^g k_i x_i \right] = -\lambda x_c \\ &= -\lambda D_c(p_c, p_{g+1}, \dots, p_n, M) \\ \frac{\partial v}{\partial p_{g+j}} &= \frac{\partial u^*}{\partial p_{g+j}} = -\lambda x_{g+j} \\ &= -\lambda D_{g+j}(p_c, p_{g+1}, \dots, p_n, M) \quad (j = 1, \dots, n-g) \end{aligned} \quad [\text{D.3}]$$

Thus the indirect utility function $v(\cdot)$ can be used in place of the indirect utility function $u^*(\cdot)$, and the demand functions depend on the price index, rather than the individual prices p_1, \dots, p_g .

Separability

The composite commodity theorem tells us that we can group commodities together on the basis of a property of their relative prices. Knowing conditions under which it is possible to group commodities is important for empirical demand analysis, because data typically only exist for aggregates of commodities – food, clothing, transport, etc. – rather than for individual commodities in the sense of the theory. Unfortunately, it is often unreasonable to assume that the relative prices of the components of these aggregates have remained constant and so the composite commodity theorem cannot be applied. In such cases reliance

is placed on a restriction on the form of the utility function, usually some kind of *separability* assumption. Here we consider two such assumptions: *weak separability* and *additive separability*.

Under weak separability the n commodities can be sorted into sub-groups, denoted by vectors x^k , $k = 1, \dots, K$, in such a way that the preference ordering over the goods in one sub-group is independent of the quantities of goods in another sub-group. Another way of putting this is to say that the marginal rate of substitution between two goods in one sub-group is independent of the quantities of other goods in other subgroups. The utility function

$$u = u[v^1(x^1), v^2(x^2), \dots, v^K(x^K)] \quad [D.4]$$

expresses the idea of weak separability exactly. We have

$$MRS_{ij} = \frac{(\partial u / \partial v^k)(v_i^k)}{(\partial u / \partial v^k)(v_j^k)} = \frac{v_i^k(x^k)}{v_j^k(x^k)} \quad k = 1, \dots, K \quad [D.5]$$

if goods i and j are in the same sub-group.

If we were to solve the problem of maximizing u subject to the budget constraint, we would find that we had K subsets of conditions of the form $v_i^k / v_j^k = p_i / p_j$, $k = 1, \dots, K$, where the left-hand sides of these equations depend only on the quantities of goods in the k th sub-group and p_i is the price of the i th good in that sub-group. If we knew the consumer's optimal amount of expenditure on each sub-group, say M_k , where $\sum_{k=1}^K M_k = M$, then we could solve separately for the demand functions of each sub-group and they could be written

$$x_i = D_i^k(p^k, M_k) \quad k = 1, \dots, K \quad [D.6]$$

that is as functions *only* of the vector of prices of the goods in the sub-group, p^k , and expenditure on that sub-group.

We could only find the M_k from the full solution to the consumer's problem but it is useful to know that the consumer's demand functions take the form [D.6]. We can then think of the consumer as first allocating optimally the expenditures M_k to each sub-group of goods, and then obtaining demands for the individual commodities by solving the problem

$$\max v^k(x^k) \text{ s.t. } p^k x^k = M_k \quad k = 1, \dots, K \quad [D.7]$$

This can be handled theoretically in the following way. From [D.7] we will have the K indirect utility functions $\phi^k(p^k, M^k)$ giving the overall indirect utility function

$$u = u^*(\phi^1(p^1, M_1), \dots, \phi^K(p^K, M_K))$$

With the prices held constant, we can solve the problem for the optimal M_k

$$\max u \text{ s.t. } \sum_{k=1}^K M_k = M$$

which tells us that at the optimal expenditure allocation

$$\frac{\partial u^*}{\partial \phi^k} \frac{\partial \phi^k}{\partial M_k} = \lambda \quad k = 1, \dots, K \quad [D.8]$$

where λ is the Lagrange multiplier attached to the constraint in [D.8] and so is also the consumer's marginal utility of income. Thus expenditure is optimally allocated when the marginal utilities of expenditure allocated to each sub-group are equal. Inserting the optimal expenditures into the indirect utility functions ϕ^k and applying Roy's identity gives us the individual commodity demands.

When preferences exhibit additive separability the form of the utility function is

$$u = F[u_1(x_1) + u_2(x_2) + \dots + u_n(x_n)] \quad F'[\cdot] > 0 \quad [D.9]$$

i.e. any positive monotonic transformation of a sum of individual utility functions. This functional form has a long history in economics, and underlies the cardinal utility-based demand theory of Alfred Marshall. However, it has some rather undesirable implications. In particular, it can be shown that it rules out the existence of goods which are Hicksian complements and goods which are inferior (and so it also rules out Giffen goods). (See Question 4, Exercise 4D.) Nevertheless, two of the most widely used forms of utility function, $u = x_1^{a_1} x_2^{a_2} \dots x_n^{a_n}$ and $u = (x_1 - c_1)^{a_1} (x_2 - c_2)^{a_2} \dots (x_n - c_n)^{a_n}$ are of this form since we have the transformations

$$\log u = \sum_i a_i \log x_i; \quad \log u = \sum_i a_i \log (x_i - c_i).$$

There are many types of separability, and the analysis of their full implications and interrelationships is a large topic. The interested reader should follow up the references given at the end of this chapter.

Homotheticity

A homothetic utility function takes the form

$$u = T[f(x_1, \dots, x_n)] \quad T' > 0 \quad [D.10]$$

where f is a *linear homogeneous* function. That is, a homothetic function is a positive monotonic transformation of a linear homogeneous function. In Chapter 9 we study the properties of linear homogeneous production functions in some depth and will not duplicate the discussion here. In the case of a utility function it is clearly not permissible to restrict attention to the linear homogeneous case because utility is not cardinally measurable. It makes sense to say that doubling inputs always doubles outputs, while it does not make sense to say doubling consumption quantities always doubles utility, since the utility function can always be transformed in such a way as to make this statement false.

However, we can exploit one parallel. In Chapter 9 it is shown that if the production function $y = f(x_1, \dots, x_n)$ is linear homogeneous then the cost function $C = C(w_1, \dots, w_n, y)$, where the w_i are the input prices, takes the form $C = c(w_1, \dots, w_n)y$. That is, it can be factored into a unit cost function of input prices alone, and output. Now the cost minimization problem for the firm which gives that result is identical in structure to the expenditure minimization problem.

$$\min \sum p_i x_i \text{ s.t. } f(x_1, \dots, x_n) = u \quad [D.11]$$

where we choose f to be linear homogeneous. Thus, in this case we can write the expenditure function as

$$m(p_1, \dots, p_n, u) = a(p_1, \dots, p_n)u \quad [\text{D.12}]$$

Now, transforming the utility function f in [D.11] by some positive monotonic transformation $T[\cdot]$ cannot change the solution vector x^* for the problem, and hence the expenditure value $\sum p_i x_i^*$. It simply changes the value of u in the constraint and cannot alter the form of the function in [D.12]. That is, simply relabelling the consumer's indifference curves with a different set of numbers does not change the expenditure required to reach any specified indifference curve.

The expenditure function [D.12] has some very strong implications for the demand functions. Inverting $m(p, u)$ to get the indirect utility function $u^*(p, M)$ yields

$$u^*(p, M) = M/a(p) \quad [\text{D.13}]$$

Applying Roy's identity to get the Marshallian demand functions we have

$$D_i(p, M) = \frac{-\partial u^*/\partial p_i}{\partial u^*/\partial M} = \frac{M}{a} \frac{\partial a}{\partial p_i} = M \frac{a_i(p)}{a(p)} \quad [\text{D.14}]$$

so that demand for good i is proportional to income and the Engel curve, plotting consumption against income is a straight line through the origin. Since [D.14] implies that $\log x_i = \log M + \log(a_i/a)$ we see that the income elasticity of demand for good i is

$$\eta_i = \frac{\partial x_i/x_i}{\partial M/M} = \frac{\partial \log x_i}{\partial \log M} = 1 \quad i = 1, \dots, n \quad [\text{D.15}]$$

The expenditure or budget share $s_i = p_i x_i/M$ is also independent of income so that the consumer always spends a constant proportion of income on a commodity as income varies.

Quasi-homothetic preferences, due to W. M. Gorman, give an expenditure function of the form

$$m(p, u) = a(p_1, \dots, p_n) + b(p_1, \dots, p_n)u \quad [\text{D.16}]$$

where a could be interpreted as a level of expenditure required for 'subsistence' ($u = 0$). Setting $M = m(p, u)$, invert the expenditure function to get the indirect utility function

$$u = (M - a)/b \quad [\text{D.17}]$$

Using Shephard's lemma in [D.16] and substituting from [D.17] gives

$$x_i = \frac{\partial a}{\partial p_i} + \frac{(M - a)}{b} \frac{\partial b}{\partial p_i} \quad [\text{D.18}]$$

(Alternatively use Roy's identity on [D.1]). Thus, for given prices, the Engel curve relating x_i and M is again a straight line, but no longer a ray through the origin. The expenditure share $p_i x_i/M$ is no longer constant and expenditure elasticities of demand are no longer identical and equal to unity.

Exercise 4D

1. In the treatment of the composite commodity theorem, express the requirement that relative prices for a group of commodities remain unchanged by setting $p_i = k p_i^0$, $i = 1, \dots, g$, where $k > 0$ is the same for all i but can itself vary, and p_i^0 is some constant base price. Show that the composite commodity theorem continues to hold, with k taken as the price of the composite commodity. Derive the expenditure function and show that it has the properties set out in section A, with k as the price of the composite commodity.
2. Consider the utility function $u = (\alpha_1 x_1^{-\theta} + \alpha_2 x_2^{-\theta})^{-1/\theta}$. What properties discussed in this section are true of this function?
3. Show that the Stone-Geary utility function $u = (x_1 - c_1)^\alpha (x_2 - c_2)^{1-\alpha}$, where c_1 and c_2 are subsistence consumption levels, has an expenditure function of the form [D.16].
4. Show that additive separability of the utility function rules out the possibilities that goods are (a) inferior and (b) Hicksian complements.

E. Aggregation*

Up until now we have discussed the properties of individual demand functions. The market demand function is found by summing these individual functions. Thus if $D_i^h(p_1, \dots, p_n, M^h)$ denotes the Marshallian demand function of individual h for commodity i , with M^h as h 's income, the market Marshallian demand function is

$$D_i(p_1, \dots, p_n, M^1, \dots, M^H) = \sum_{h=1}^H D_i^h(p_1, \dots, p_n, M^h) \quad [\text{E.1}]$$

In general, the market demand is a function of the entire set of individual incomes. The effect of a change in aggregate income $M = \sum_{h=1}^H M^h$ on market demand depends on the consumers' income elasticities of demand and on the way in which the income is distributed among them. It is then of interest to ask under what conditions it would be possible to write the market demand as a function of prices and aggregate income only, i.e. $D_i(p_1, \dots, p_n, M)$. This is the way that market demand functions are usually written in elementary supply and demand analysis, but we shall now see that the conditions under which this is correct are in fact very stringent.

If a change in the distribution of income, with its total fixed, is not to change market demand, the effect of taking £1 away from consumer i and giving it to consumer j must be zero. But this requires that the reduction in consumer i 's demand be equal to the increase in consumer j 's demand (or conversely for an inferior good). This must be true for all possible pairs of consumers and at all possible income levels. The individual demand functions must therefore be linear in income, with the same coefficients on income for all consumers:

$$x_i^h = \alpha_i^h(p_1, \dots, p_n) + \beta_i^h(p_1, \dots, p_n)M^h, \quad h = 1, \dots, H \quad [\text{E.2}]$$

so that aggregating gives market demand as

$$x_i = \sum_h x_i^h = \sum_h \alpha_i^h(p_1, \dots, p_n) + \beta_i(p_1, \dots, p_n)M \quad [\text{E.3}]$$

Although both coefficients α_i^h and β_i can be functions of prices and the α_i^h can vary across households, the income coefficient β_i must be the same for all households. (Note that this does not imply that income elasticities must be the same.)

The requirement [E.2] for aggregation restricts the consumer's preferences severely since all consumers must have linear Engel curves with identical slopes. We will show that *Marshallian demands satisfy [E.2] if and only if all consumers have quasi-homothetic preferences with similar expenditure functions: $m^h = a^h(p) + b(p)u^h$* . (Here u^h is the utility function of consumer h .) Note that the expenditure functions can differ across individuals only in the term $a^h(p)$. To establish the sufficiency of the assertion recall from [D.18] that if preferences are quasi-homothetic the Marshallian demands can be written as

$$x_i^h = \alpha_i^h(p) + \frac{(M^h - a^h(p))}{b^h(p)} \beta_i^h(p) = \left[\alpha_i^h(p) - \frac{a^h(p)}{b^h(p)} \beta_i^h(p) \right] + \frac{\beta_i^h(p)}{b^h(p)} M^h \quad [\text{E.4}]$$

The first bracketed term on the right-hand side depends only on p and so does the coefficient on M^h . Thus if $b^h(p)$ is the same function for all individuals [E.4] is of the same form as [E.2] and the individual demand functions aggregate to give the market demand as a function of total income.

To prove the necessity of quasi-homotheticity for aggregation, recall that if we set the required level of utility in the expenditure minimization problem equal to the maximized utility achieved in the utility maximization problem, the Marshallian and Hicksian demands coincide:

$$D_i(p, M) = H_i(p, u^*(p, M)) \quad [\text{E.5}]$$

(we drop the use of superscripts to identify individuals for the moment). Differentiating [E.5] with respect to M gives

$$D_{iM}(p, M) = H_{iu}(p, u^*)u_M^*(p, M) = H_{iu}(p, u^*)\lambda(p, M) \quad [\text{E.6}]$$

where $D_{iM} = \partial D_i / \partial M$, $H_{iu} = \partial H_i / \partial u$ and $u_M^* = \partial u^* / \partial M = \lambda$ (Remember the Lagrange multiplier in the utility maximization problem is the marginal utility of money income.) Now [E.2] requires that the second derivative of D_i with respect to M is zero and so differentiating [E.6]

$$D_{iMM}(p, M) = H_{iui}(p, u^*)[\lambda(p, M)]^2 + H_{iu}(p, u^*)\lambda_M(p, M) = 0 \quad [\text{E.7}]$$

where the subscripts on D_i , H_i now denote second order partial derivatives and $\lambda_M = \partial \lambda / \partial M$ is the rate of change of the marginal utility of money income. With prices held constant it is always possible to choose the utility function $u(x)$ so that the marginal utility of money income is constant with respect to M : $\lambda_M = 0$. If we do so then [E.7] restricts the consumer's Hicksian demand functions to satisfy

$$H_{iui}(p, u^*) = 0 \quad [\text{E.8}]$$

Integrating [E.8] with respect to u , holding p constant, implies

$$H_{iu} = \frac{\partial^2 m(p, u)}{\partial p_i \partial u} = g_i(p) \quad [\text{E.9}]$$

and integrating again with respect to u gives

$$H_i(p, u) = \frac{\partial m(p, u)}{\partial p_i} = f_i(p) + g_i(p)u \quad [\text{E.10}]$$

Multiplying the Hicksian demands by p_i and summing over i yields the expenditure function

$$\sum_i p_i H_i(p, u) = m(p, u) = \sum_i p_i f_i(p) + \sum_i p_i g_i(p)u \quad [\text{E.11}]$$

Since the first term on the right-hand side depends only on p and can be replaced with $a(p)$ and the coefficient on u can be replaced with $b(p)$, [E.11] is the expenditure function for an individual with quasi-homothetic preferences. Thus, reverting now to superscripts to identify individuals, we have shown that Engel curves are straight lines, as required by [E.2], only if the expenditure function for individual h can be written as $m^h = a^h(p) + b^h(p)u^h$. The requirement in [E.2] that the Engel curves have the same slope implies further that $b^h(p)$ must be the same for all h . Hence we have established that aggregation requires that individuals have expenditure functions $m^h = a^h(p) + b(p)u^h$.

Note that if this condition is satisfied [E.3] could be regarded as the Marshallian demand function of a single 'aggregate' consumer who chooses the market demand x_i optimally subject to a budget constraint determined by the prices p_i and aggregate market income M . This consumer could be thought of as possessing quasi-homothetic preferences with the expenditure function

$$m = a + bu$$

where the functions a and b satisfy

$$\sum_i \alpha_i^h = \frac{\partial a}{\partial p_i} - \frac{a}{b} \frac{\partial b}{\partial p_i}; \quad \beta_i = \frac{1}{b} \frac{\partial b}{\partial p_i} \quad [\text{E.12}]$$

and $\sum \alpha_i^h$ and β_i would be observed from market data.

Finally, we can consider whether anything about the aggregate Hicksian market demand functions could be inferred from estimates of the parameters of an aggregate Marshallian demand function (recall the discussion for the case of individual demand functions in section C). The assumption so far made is not strong enough: if the α_i^h differ across consumers then nothing can be inferred about the constant terms of the individual expenditure functions from $\sum \alpha_i^h$. However, if the α_i^h are all identical, then we can use [E.4] to identify from market data the parameters of the individual expenditure functions. We can then obtain aggregate compensating and equivalent variation measures. (Note, we are implying a particular value judgement about income distribution when we sum these across consumers, namely that society regards the social utility of the marginal £1 income to each consumer as the same.) The assumption that consumers have identical quasi-homothetic preferences, implying identical linear Engel curves for all goods, is a very strong one.

Notes

1. Young's Theorem: if a function of n variables $f(x)$ has continuous second-order partial derivatives, then the cross-partial derivatives are equal: $f_{ij}(x) = f_{ji}(x)$.

References and further reading

For an excellent treatment of the entire field of consumer theory, including empirical estimation of demand systems, see

A. Deaton and J. Muellbauer. *Economics and Consumer Behaviour*, Cambridge University Press, Cambridge, 1980.

For a succinct survey, see

R. Blundell. 'Consumer behaviour: theory and empirical evidence – a survey', *Economic Journal*, 98, 1988, 16–65.

On duality see

W. E. Diewert. 'Applications of duality theory', in *Frontiers of Quantitative Economics*, vol II, M. D. Intriligator and J. W. Kendrick (eds), North Holland, Amsterdam, 1974.

A thorough analysis of functional forms is given in

C. Blackorby, D. Primont and R. R. Russell. *Duality, Separability and Functional Structure*. American Elsevier, New York, 1978.

For accounts of the integrability problem see

P. A. Samuelson. 'The problem of integrability in utility theory', *Economica*, 17, 1950, 355–85.
L. Hurwicz and H. Uzawa. 'On the integrability of demand functions', in *Preference, Utility and Demand*, J. S. Chipman, L. Hurwicz, M. K. Richter and H. F. Sonnenschein (eds), Harcourt Brace, New York, 1971.

The various measures of consumer surplus are compared in

P. A. Samuelson. *Foundations of Economic Analysis*. Harvard University Press, Harvard, 1947, 189–202.

J. S. Chipman and J. C. Moore. 'Compensating variation, consumer surplus, and welfare', *American Economic Review*, 70, 1980, 933–49.

G. W. McKenzie. *Measuring Economic Welfare: New Methods*, Cambridge University Press, Cambridge, 1983.

J. L. Wrigglesworth and H. S. E. Gravelle. 'The three consumer surpluses as individual welfare measures', *Scottish Journal of Political Economy*, 34, 1987, 230–48.

and above all in

J. R. Hicks. 'The rehabilitation of consumer's surplus', *Review of Economic Studies*, VIII, 1940–1, 108–16.

CHAPTER 5

Further models of consumer behaviour*

A. Revealed preference

We emphasized in Chapter 3 that utility functions are convenient numerical representations of preferences and that neither they nor the consumer's preferences are directly observable. This subjectivity of the foundations of consumer theory stimulated interest in the development of a theory of demand based solely on observable and measurable phenomena, namely the bundles actually bought by a consumer and the prices and money incomes at which they were bought. The emphasis in this approach is on assumptions about the consumer's behaviour, which can be observed, rather than on his preferences, which cannot.

As in the utility theory of Chapter 3, we assume that the consumer faces a given price vector, p , and has a fixed money income, M . Our first behavioural assumption is that the consumer spends all his income, which has similar implications to assumption 4 of section 3A.

The second assumption is that only one commodity bundle x is chosen by the consumer for each price and income situation. In other words, confronted by a particular p vector and having a particular M , the consumer will always choose the same bundle.

The third assumption is that there exists one and only one price and income combination at which each bundle is chosen. It is assumed, therefore, that for a given x there is some p, M situation in which x will be chosen by the consumer and that situation is unique.

The fourth and crucial assumption is that the consumer's choices are consistent. By this we mean that, if a bundle x^0 is chosen and a different bundle x^1 could have been chosen, then when x^1 is chosen x^0 must no longer be a feasible alternative.

To amplify this, let p^0 be the price vector at which x^0 is chosen. Then if x^1 could have been chosen when x^0 was actually chosen, the cost of x^1 , p^0x^1 , must be no greater than the cost of x^0 , which is p^0x^0 . This latter is also the consumer's money income $M_0 = p^0x^0$ when x^0 is chosen.

Similarly, let p^1 be the price vector at which x^1 is chosen. Then x^0 could not have been available at prices p^1 , otherwise it would have been chosen. That is, its cost p^1x^0 must exceed the cost of x^1 , p^1x^1 , which equals the consumer's money income M_1 when x^1 is chosen. Hence this fourth assumption can be stated succinctly as

$$p^0x^0 \geq p^0x^1 \quad \text{implies} \quad p^1x^1 < p^1x^0 \quad [A.1]$$

when x^0 is chosen at p^0, M_0 and x^1 at p^1, M_1 . If x^0 is chosen when x^1 is purchasable x^0 is said to be *revealed preferred* to x^1 . The statement [A.1] is usually referred to as the *weak axiom of revealed preference*.

This set of mild behavioural assumptions will generate all the utility based predictions of section 3D concerning the consumer's demand functions. Consider first the sign of the substitution effect. Figure 5.1 shows the consumer's initial budget line B_0 , defined by price vector p^0 and money income M_0 . The bundle chosen initially on B_0 is x^0 . B_1 is the budget line after a fall in p_1 with M unchanged, and x^1 the new bundle chosen on B_1 . Our behavioural assumptions do not place any restrictions on the location of x^1 on B_1 (explain). (Neither do the preference assumptions of section 3A, as section 3D shows.) As in section 3D, it is useful to partition the price effect (x^0 to x^1) into a change in x due solely to relative price changes (the substitution effect) and a change due solely to a change in real income. Since we have forsworn the use of utility functions in this section we cannot use the indifference curve through x^0 to define a constant real income. Instead we adopt the constant purchasing power or Slutsky definition of constant real income (see section 3D). Accordingly, the consumer's money income is lowered until, facing the new prices, he is just able to buy the initial bundle x^0 . In Fig. 5.1 the budget line is shifted inward parallel with B_1 , until at B_2 it passes through x^0 . The consumer confronted with B_2 will buy the bundle x^2 to the right of x^0 . Therefore x^0 to x^2 is the substitution effect and x^2 to x^1 the income effect of the fall in p_1 .

We can now prove that if the consumer satisfies assumption [A.1] the substitution effect must always lead to an increase in consumption of the good whose price has fallen. This is easily done in the two-good example of Fig. 5.1. x^2 must lie on B_2 (by the assumption that all income is spent) and hence there are three possibilities: x^2 can be to the left or the right of, or equal to, x^0 . x^2 cannot be to the left of x^0 on B_2 because these bundles are inside the consumer's initial feasible set and were rejected in favour of x^0 . x^2 cannot equal x^0 because the prices at which x^2 and x^0 are chosen differ and, by our second assumption, different bundles are chosen in different price-income situations. Therefore x^2 must contain more x_1 than (i.e. be to the right of) x^0 .

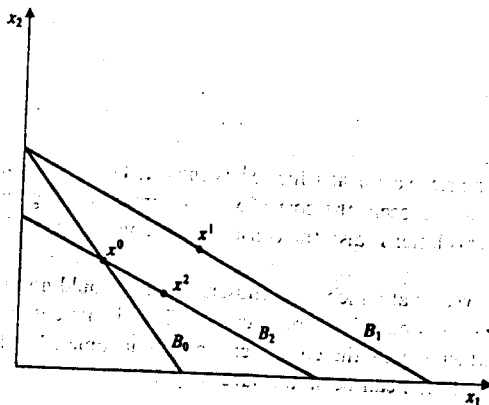


Fig. 5.1

This result can be extended to the n -good case, and the proof is instructive because similar arguments will be used in section 9D to derive comparative statics predictions in the theory of the firm. We can generalize the steps in the analysis of Fig. 5.1 as follows. p^0, x^0 are the initial price vector and consumption bundle, p^1 and x^1 are the new price vector and consumption bundle. The consumer's income is adjusted until at M_2 he can just purchase x^0 at the new prices, p^1 , so that $p^1 x^0 = M_2$. Faced with price vector p^1 and the compensated money income, M_2 , the consumer chooses x^2 and because he spends all his money income we have that $p^1 x^2 = M_2$. Hence the compensating change in M ensures that

$$p^1 x^0 = M_2 = p^1 x^2 \quad [A.2]$$

Now x^2 is chosen when x^0 is still available (i.e. they are both on the same budget plane) so that by our consistency assumption [A.1] we have

$$p^0 x^0 < p^0 x^2 \quad [A.3]$$

or: x^2 was not purchasable when x^0 was bought. Rearranging [A.2] gives

$$p^1 x^0 - p^1 x^2 = p^1 (x^0 - x^2) = 0 \quad [A.4]$$

and similarly [A.3] gives

$$p^0 x^0 - p^0 x^2 = p^0 (x^0 - x^2) < 0 \quad [A.5]$$

Subtracting [A.5] from [A.4] gives

$$p^1 (x^0 - x^2) - p^0 (x^0 - x^2) = (p^1 - p^0)(x^0 - x^2) > 0$$

and multiplying by (-1) we have

$$(p^1 - p^0)(x^2 - x^0) < 0 \quad [A.6]$$

This prediction applies irrespective of the number and direction of price changes, but in the case of a change in the j th price only, p^1 and p^0 differ only in p_j and so [A.6] becomes

$$\sum_i (p_i^1 - p_i^0)(x_i^2 - x_i^0) = (p_j^1 - p_j^0)(x_j^2 - x_j^0) < 0 \quad [A.7]$$

Hence when p_j changes the substitution effect ($x_j^2 - x_j^0$) is of opposite sign to the price change. The constant purchasing power demand curve will therefore slope downwards.

We can also derive the Slutsky equation of section 4B from the behavioural assumptions. Since $M_2 = p^1 x^0$ and $M_0 = p^0 x^0$ the compensating reduction in M is

$$\Delta M = M_0 - M_2 = p^0 x^0 - p^1 x^0 = (p^0 - p^1)x^0 = -(p^1 - p^0)x^0$$

and in the case of a change (Δp_i) in p_i only we have

$$\Delta M = -\Delta p_i x_i^0 \quad [A.8]$$

The price effect of p_i on x_j is $(x_j^1 - x_j^0)$ and this can be partitioned into the substitution ($x_j^2 - x_j^0$) and income ($x_j^1 - x_j^2$) effects:

$$x_j^1 - x_j^0 = (x_j^2 - x_j^0) + (x_j^1 - x_j^2)$$

Dividing this by Δp_i gives

$$\frac{x_j^1 - x_j^0}{\Delta p_i} = \frac{x_j^2 - x_j^0}{\Delta p_i} + \frac{x_j^1 - x_j^2}{\Delta p_i} \quad [\text{A.9}]$$

But from [A.8] $\Delta p_i = -\Delta M/x_i^0$ and substituting this in the second term on the right-hand side of [A.9] yields

$$\begin{aligned} \frac{x_j^1 - x_j^0}{\Delta p_i} &= \frac{x_j^2 - x_j^0}{\Delta p_i} - x_i^0 \frac{(x_j^1 - x_j^2)}{\Delta M} \\ \frac{\Delta x_j}{\Delta p_i}|_M &= \frac{\Delta x_j}{\Delta p_i}|_{px} - x_i^0 \frac{\Delta x_j}{\Delta M}|_p \end{aligned} \quad [\text{A.10}]$$

The $|_M$ notation indicates that money income is held constant in evaluating the rate of change of x_j with respect to p_i , and the similar notation on the righthand side that purchasing power px and price vector p are being held constant in evaluating the rate of change of x_j with respect to p_i and M . [A.10] is the discrete purchasing power version of the Slutsky equation of section 4B.

It is possible to show that the utility maximizing theory of the consumer and the revealed preference theory are equivalent: all the predictions derived from the assumption about preferences in section 3A can also be derived from the assumption about behaviour made in this section. A consumer who satisfies the preference assumptions will also satisfy these behavioural assumptions. Similarly, if the consumer satisfies the behavioural assumptions, from his choices we can construct curves which have all the properties of the indifference curves of section 3A. And so he can be thought of as acting as if he possessed preferences satisfying the preference assumptions. (Strictly the weak axiom needs to be strengthened slightly.) Since the two theories are equivalent we will not consider any more of the predictions of the theory of revealed preference but will instead use the theory to investigate some properties of price indices.

Price indices

As we noted in section 4C, it is often useful to be able to measure the benefits to consumers of changes in prices of goods. For example, a government may wish to pay state pensions which ensure at least a constant level of utility to its pensioners in a period when most prices of goods bought by pensioners fluctuate. The pensions, i.e. money incomes, must therefore be adjusted as prices vary, but by how much?

Let x^0, x^1 be the bundles of goods bought by a consumer with incomes M_0, M_1 at price vectors p^0, p^1 respectively. (So that $p^0 x^0 = M_0$ and $p^1 x^1 = M_1$ and 0 denotes the initial or base period and 1 the current period.) Suppose the consumer satisfies our behavioural assumptions (or equivalently the preference assumptions of section 3A). Under what circumstances can we say that he is better off in one price-income situation than another?

Suppose first that

$$p^1 x^1 \geq p^1 x^0 \quad [\text{A.11}]$$

so that x^1 is revealed preferred to x^0 , in that x^1 was chosen when x^0 was available. Dividing

both sides of [A.11] by $p^0 x^0$ gives

$$MI = \frac{p^1 x^1}{p^0 x^0} \geq \frac{p^1 x^0}{p^0 x^0} = LP \quad [\text{A.12}]$$

The left-hand side of [A.12] is an index of the consumer's money income and the right-hand side is an index of prices with base period quantities as weights, known as the *Laspeyres* price index. Hence, if the money income index is at least as large as the Laspeyres price index the consumer will be better off. Note that if the inequality in [A.12] was $<$ rather than \geq nothing could be inferred from the relationship of the two indices.

Now assume that

$$p^0 x^0 \geq p^0 x^1 \quad [\text{A.13}]$$

so that x^0 is revealed preferred to x^1 . [A.13] is equivalent to

$$\frac{1}{p^0 x^0} \leq \frac{1}{p^0 x^1}$$

and hence to

$$MI = \frac{p^1 x^1}{p^0 x^0} \leq \frac{p^1 x^1}{p^0 x^1} = PP \quad [\text{A.14}]$$

where PP is the *Paasche* current weighted price index. If the money income index is less than the Paasche price index the consumer is definitely worse off in the current period than in the base period. Again if $<$ replaces \geq in [A.13] (so that $>$ replaces \leq in [A.14]) nothing can be said about whether the individual is better or worse off.

In some circumstances therefore comparisons of price and money income indices do tell us whether a consumer is better or worse off as a result of changes in prices and his income, without requiring detailed information on his preferences.

Price indices are not, however, calculated for each individual using his consumption levels as weights. The weights used are either total or average consumption bundles for particular groups (e.g. all pensioners, or the inhabitants of particular regions). Suppose that the Laspeyres price index and the money income index are calculated using the sum of consumption bundles and money incomes:

$$MI = \frac{\sum M_1}{\sum M_0} = \frac{\sum p^1 x^1}{\sum p^0 x^0} = \frac{p^1 \sum x^1}{p^0 \sum x^0} \quad [\text{A.15}]$$

$$LP = \frac{p^1 \sum x^0}{p^0 \sum x^0} \quad [\text{A.16}]$$

where M_0^i, x^0, M_1^i, x^1 are the bundle and income of individuals in the base and current periods. What can be inferred from the relationship between [A.15] and [A.16]? Assume that MI exceeds LP and multiply both indices by $p^0 \sum x^0$ to give

$$p^1 \sum x^1 > p^1 \sum x^0 \quad [\text{A.17}]$$

which, taking a case involving two consumers, a and b , for simplicity, can be written

$$p^1 x^{a1} + p^1 x^{b1} > p^1 x^{a0} + p^1 x^{b0} \quad [\text{A.18}]$$

Now [A.18] does *not* imply that $p^1 x^{a1} > p^1 x^{a0}$ and $p^1 x^{b1} > p^1 x^{b0}$, but merely that *at least one* of these inequalities holds, so that at least one of the consumers is better off in the current period. It is possible, however, that one of the consumers may be worse off. Hence $MI > LP$ does not imply that *all* members of the group for whom the indices are calculated are better off, merely that *some* of them are.

In some circumstances [A.18] will imply that *a* and *b* are better off in the current period. Suppose that the bundles bought by the consumers at given prices are proportional, i.e. that $x^{a1} = kx^{b1}$ and $x^{a0} = kx^{b0}$. Hence [A.18] is equivalent to

$$(1+k)p^1 x^{b1} > (1+k)p^1 x^{b0} \quad [\text{A.19}]$$

and so

$$p^1 x^{b1} > p^1 x^{b0} \quad [\text{A.20}]$$

so that consumer *b* is better off. But multiplying both sides of [A.20] by *k* gives

$$p^1 kx^{b1} = p^1 x^{a1} > p^1 x^{a0} = p^1 kx^{b0}$$

and consumer *a* is better off as well. If the consumers in a group have preferences which ensure that each spends the same proportion of their income on the same good then price and money income indices can tell us, for some price and income changes, whether *all* consumers in the group are better or worse off. In order for the consumers to have equal proportionate expenditure patterns for all price vectors one of two conditions must be satisfied:

- consumers have identical preferences and identical incomes so that they buy identical bundles ($k = 1$ in the above example);
- consumers have identical *homothetic* preferences so that income consumption curves (see section 4D) are straight lines from the origin. Each good will have the same proportion of the consumer's income spent on it irrespective of the size of his income and income elasticities of demand for all goods will be unity.

The group of consumers for whom the indices are calculated must satisfy one of the above conditions if the indices are to be of use. This suggests that there may need to be many such indices and that the indices should be frequently updated. This latter suggestion implies that the periods being compared should be not too far apart, in order to minimize the errors from non-unitary income elasticities which can arise if incomes differ even though groups have identical tastes.

Exercise 5A

- Show that a consumer who satisfies the preference assumptions of section 3A will also satisfy the behavioural assumptions. Can you relate the assumptions in the two sections? Which behavioural assumption, for example, plays a similar role to the transitivity assumption of section 3A?

- Draw diagrams to show that $MI < LP$ and $MI > PP$ tell us nothing about which situation is preferred.
- Suppose that the actual weights used in a price index are average consumption bundles for the group of consumers. Under what conditions does $MI > LP$ imply that *all* consumers are now better off?
- Do the remarks in the last part of the section and the results obtained in Question 3 hold for Paasche price indices?
- * Laspeyres and Paasche quantity indices have the form

$$LQ = \frac{p^0 x^1}{p^0 x^0} \quad PQ = \frac{p^1 x^1}{p^1 x^0}$$

If $LQ \geq 1$ or $PQ \leq 1$ can anything be said about whether the individual consuming x^0 and x^1 is better or worse off? Suppose the quantities were the total consumption of all members of an economy. Could anything be said about changes in standards of living using the indices?

- Suppose that the government increases the income of its pensioners in proportion to the rise in the Laspeyres price index. Will they be better or worse off? What if the government used a Paasche price index? What if prices fell?

B. Consumption technology

In this section we outline an alternative approach to the standard theory of consumer behaviour, developed by K. Lancaster. The conventional theory is adequate for many purposes but there are some phenomena which can only be incorporated into the theory with great difficulty. Consider, for example, the introduction of a new good. One method of handling this would be to assume that when there are n goods the consumer has a preference ordering defined with respect to those n goods and on the introduction of a new good his preference ordering is now defined with respect to the $n+1$ goods. This procedure does not tell us how the two preference orderings are related and hence what the effect of the introduction of the new good will be on the demand for the other goods. Alternatively, we could assume that the consumer has a preference ordering defined with respect to *all* goods: those which exist now *and* those which will exist in the future. The introduction of a new good then corresponds to the reduction of the price of that good from an arbitrarily high level. This procedure is unsatisfactory in that it assumes a rather large amount of knowledge on the part of the consumer.

A second problem concerns the analytical treatment of advertising. In the standard theory, if advertising changes a consumer's demand it must do so by changing her preferences in some way. This then means that it is impossible to say whether she is better off as a result of the advertising, since we can only make welfare comparisons for a consumer with reference to a fixed set of preferences. This in turn means that we cannot discuss the

resource allocation implications of advertising – whether, for example, there is ‘too much’ advertising.

Finally, it would be interesting to have a more fundamental explanation of *why* goods are substitutes or complements for each other. In the standard theory, this is taken simply as given by the consumer’s preferences rather than an inherent property of the goods themselves that could be the subject of a deeper analysis.

Lancaster’s approach is to regard a unit of any good as a given bundle of attributes or characteristics. For example, a particular type of food will consist of specific flavours, calories, vitamins and so on. A combination of types of food will produce a particular vector of quantities of these characteristics. The consumer’s preferences are defined over bundles of characteristics and the demand for goods is then a *derived demand*: goods are in effect inputs into a production process. Consumption is the activity of extracting characteristics from goods.

This approach is very suitable for dealing with the three difficulties of the standard theory just outlined. A new good can be defined as a new way of bundling a *given* set of characteristics, and so it can easily be placed into the preference ordering because this is over characteristics (obviously Lancaster’s theory would face the same difficulty in dealing with new characteristics as the standard theory has with new goods, but the basic assumption is that characteristics are much more stable and unvarying than goods). Advertising can be regarded as the transmission of information (which may or may not be correct and complete) about the characteristics of goods, and Lancaster’s formulation allows us also to model in a tractable way uncertainty about preferences for a good, in terms of the uncertainty about the characteristics a good actually possesses. Finally, the reason goods are complements or substitutes lies in the nature of the characteristics they possess, and can be given expression in those terms.

To put the theory more formally, let $a = (a_1, \dots, a_r)$ be the vector of all the possible attributes that goods may possess, and $x = (x_1, \dots, x_n)$ the vector of all goods. Then, one unit of good x_j yields a vector $\alpha_j = (\alpha_{1j}, \dots, \alpha_{rj})$ of the attributes, where each $\alpha_{ij} \geq 0$ (of course many components of the vector may be zero). It is convenient to make a *linearity assumption*: α_{ij} , the amount of attribute i yielded by one unit of good j , is fixed regardless of the level of consumption of this or any other good. This assumption is largely for convenience, but note two underlying rather more substantive assumptions: the characteristics are *numerically* measurable, α_j is a vector of real numbers; and all consumers would assign the same vector α_j to a unit of good j , so attributes are *objectively* measurable and fully known. These two assumptions are strong, but they ensure a simple and powerful analysis.

Given the linearity assumption, the amount of attribute i the consumer obtains from any consumption vector x is

$$a_i = \alpha_{i1}x_1 + \dots + \alpha_{in}x_n = \sum_j \alpha_{ij}x_j \quad i = 1, \dots, r \quad [\text{B.1}]$$

This system of r equations defines the *consumption technology* of the model. The consumer’s preferences are defined on characteristics, and we assume that these preferences satisfy the assumptions set out in Chapter 3, so that they can be represented by the usual kind of utility function $u(a_1, \dots, a_r)$. Finally, the consumer has a given income and faces given

prices of goods. Thus she solves the following optimization problem:

$$\begin{aligned} \max_{a, x} u(a_1, \dots, a_r) \text{ s.t. } & \text{(i) } \sum_j p_j x_j \leq M \\ & \text{(ii) } a_i = \sum_j \alpha_{ij} x_j \quad i = 1, \dots, r \\ & \text{(iii) } x_j \geq 0 \quad j = 1, \dots, n \end{aligned} \quad [\text{B.2}]$$

An obvious way to solve this problem is to use (ii) to substitute for the a_i into the utility function and then solve the problem in ‘goods space’. This is not, however, particularly insightful. It is far more interesting to analyse the model in ‘characteristics space’. It is in characteristics space that we can best deal with the types of issues, discussed at the beginning of this section, for which this approach is particularly well suited. We do this by first using (i) and (ii) of [B.2] to obtain a type of budget constraint in characteristics space known as an *efficiency frontier*: the upper boundary of the set of characteristics vectors a consumer can achieve by combining goods. The key point is that this is independent of preferences, and its essential properties (though not its exact location) are the same for all consumers regardless of income. Once we have investigated this efficiency frontier thoroughly, we can introduce preferences and see how the optimal choice of a bundle of characteristics determines the choice of a bundle of goods.

Efficient combinations of goods

We concentrate on a model with only two characteristics and two or three goods. Figure 5.2 has the two characteristics a_1, a_2 measured along its horizontal and vertical axes. Purchase of one unit of good 1 will produce α_{11} units of characteristic 1 and α_{21} units of characteristic 2, and so x_1 units of good 1 produce $\alpha_{11}x_1, \alpha_{21}x_1$ units of the characteristics. When only good 1 is bought, the ray OG_1 in the figure shows the combinations of a_1, a_2 produced by varying the level of x_1 . OG_2 similarly shows combinations of characteristics

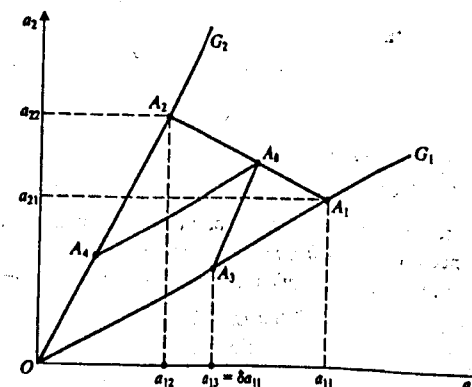


Fig. 5.2

generated by varying the level of good 2 with x_1 set equal to zero. Notice that good 2 has by assumption a higher ratio of a_2 produced to a_1 produced than does good 1.

Now suppose the consumer spends her entire income M on good 1. She will be able to buy M/p_1 units of x_1 and so produce $\alpha_{11}M/p_1 = a_{11}$, $\alpha_{21}M/p_1 = a_{21}$ units of the two characteristics. Hence if only x_1 is bought and all income spent the consumer will be at $A_1 = (a_{11}, a_{21})$ in the figure. Similarly, if M is spent only on good 2, M/p_2 units of good 2 will be bought, putting the consumer at $A_2 = (a_{12} = \alpha_{12}M/p_2, a_{22} = \alpha_{22}M/p_2)$ on OG_2 .

The assumptions about the consumption technology imply that if two bundles of characteristics A' and A'' are feasible then any *convex combination* of them; $\delta A' + (1 - \delta)A''$ ($0 \leq \delta \leq 1$) is also feasible (see section 2B). Hence all points in the area OA_1A_2 are feasible. In particular, by spending all her income in various proportions on the two goods the consumer can achieve any characteristics combination along A_1A_2 .

Conversely, given any point on this line A_1A_2 , we can find the mixture of x_1 and x_2 required to attain it. Thus consider A_δ in Fig. 5.2. By drawing a line from A_δ parallel with OG_2 back to OG_1 we get to A_3 on OG_1 . By drawing a line parallel with OG_1 from A_δ to OG_2 we get to A_4 . Recalling the 'parallelogram rule' for the addition of vectors, we see that the sum of A_3 and A_4 is A_δ . A_3 is the combination achieved by spending δM on good 1 and A_4 the combination achieved by spending $(1 - \delta)M$ on good 2.

The line A_1A_2 is then the efficiency frontier or budget line in characteristics space. Its slope defines the rate at which the consumer can trade off one characteristic for the other, by varying the quantities of the two goods she buys. Since the goods contain the two characteristics in different proportions, moving along the budget constraint in goods-space allows a movement along the efficiency frontier in characteristics space. For example, increasing consumption of good 1 by 1 unit implies p_1/p_2 units less of good 2. This implies a reduction of $\alpha_{12}(p_1/p_2)$ units of characteristic 1, and of $\alpha_{22}(p_1/p_2)$ units of characteristic 2. But 1 unit more of good 1 implies gaining α_{11} units of the first characteristic and α_{21} units of the second. Thus the *net* changes in characteristics are

$$\Delta a_1 = \alpha_{11} - \alpha_{12}(p_1/p_2); \quad \Delta a_2 = \alpha_{21} - \alpha_{22}(p_1/p_2) \quad [\text{B.3}]$$

and so the slope of the efficiency frontier A_1A_2 must be

$$\frac{\Delta a_2}{\Delta a_1} = \frac{\alpha_{21} - \alpha_{22}(p_1/p_2)}{\alpha_{11} - \alpha_{12}(p_1/p_2)} = \frac{\alpha_{21}/p_1 - \alpha_{22}/p_2}{\alpha_{11}/p_1 - \alpha_{12}/p_2} \quad [\text{B.4}]$$

This can be confirmed by referring to Fig. 5.2 and noting that, since $a_{ij} = \alpha_{ij}M/p_j$, $i, j = 1, 2$, we have

$$\frac{\Delta a_2}{\Delta a_1} = \frac{a_{21} - a_{22}}{a_{11} - a_{12}} = \frac{M(\alpha_{21}/p_1 - \alpha_{22}/p_2)}{M(\alpha_{11}/p_1 - \alpha_{12}/p_2)} \quad [\text{B.5}]$$

as required. We shall now explore the derivation of this slope more formally and, in doing so, introduce the important concept of the *implicit price* of an attribute.

Writing out the budget constraint and consumption technology for this problem gives

$$\begin{aligned} a_1 &= \alpha_{11}x_1 + \alpha_{12}x_2 \\ a_2 &= \alpha_{21}x_1 + \alpha_{22}x_2 \\ p_1x_1 + p_2x_2 &= M \end{aligned} \quad [\text{B.6}]$$

We can solve the first two equations of [B.6] for x_1 and x_2 (using Cramer's Rule) and substitute into the budget constraint to obtain

$$p_1 \left(\frac{a_1\alpha_{22} - a_2\alpha_{12}}{\alpha_{11}\alpha_{22} - \alpha_{21}\alpha_{12}} \right) + p_2 \left(\frac{a_2\alpha_{11} - a_1\alpha_{21}}{\alpha_{11}\alpha_{22} - \alpha_{21}\alpha_{12}} \right) = M \quad [\text{B.7}]$$

Rearranging the left hand side of [B.7] gives

$$\pi_1 a_1 + \pi_2 a_2 = M$$

where

$$\pi_1 \equiv \frac{p_1\alpha_{22} - p_2\alpha_{21}}{\alpha_{11}\alpha_{22} - \alpha_{21}\alpha_{12}}; \quad \pi_2 \equiv \frac{p_2\alpha_{11} - p_1\alpha_{12}}{\alpha_{11}\alpha_{22} - \alpha_{21}\alpha_{12}} \quad [\text{B.8}]$$

Then π_1 and π_2 are defined as the implicit prices of the attributes. To see why this is so, note first that their ratio is

$$\frac{\pi_1}{\pi_2} = \frac{(p_1\alpha_{22} - p_2\alpha_{21})}{(p_2\alpha_{11} - p_1\alpha_{12})} = \frac{p_1p_2(\alpha_{22}/p_2 - \alpha_{21}/p_1)}{p_1p_2(\alpha_{11}/p_1 - \alpha_{12}/p_2)} = -\frac{\Delta a_2}{\Delta a_1} \quad [\text{B.9}]$$

Thus, as usual, the ratio of prices gives the (absolute value of) the slope of the budget line. Moreover, consider the *valuation equations*

$$\begin{aligned} \pi_1\alpha_{11} + \pi_2\alpha_{21} &= p_1 \\ \pi_1\alpha_{12} + \pi_2\alpha_{22} &= p_2 \end{aligned} \quad [\text{B.10}]$$

The left-hand side of one of these equations gives the value of the bundle of characteristics contained in one unit of the good, where the valuations are made at the implicit prices. The equations say that the price of each good is just equal to the value of the characteristics a unit of it contains. Alternatively, we could think of the equations in [B.10] as allocating the value of a unit of a good – its price – between its characteristics, or *imputing* prices to characteristics which *exhaust* the value of the good. Then, if we solve the equations in [B.10], (using Cramer's Rule), we have exactly the values for π_1 and π_2 given in [B.8]. That is, π_1 and π_2 satisfy the valuation equations, which confirms their economic interpretation as implicit prices of the characteristics.

Of course, the fact that we could solve [B.8], or equivalently [B.10], for these implicit prices, is due to the equality of the number of goods and the number of characteristics. If, as would more usually be assumed, the number of goods exceeds the number of characteristics, we would appear to run into trouble. We now turn to the case of three goods and two attributes, therefore, to show that the 2×2 analysis in fact remains useful, and allows an interesting extension to the question of the introduction of a new good.

Efficient combinations with three goods

If there is a third good x_3 which produces the two characteristics, Fig. 5.3 replaces Fig. 5.2. OG_3 shows the quantities of characteristics produced by spending varying sums of money entirely on good 3. Assume that if the consumer's entire income is spent on good 3 the point $A_3 = (M\alpha_{13}/p_3, M\alpha_{23}/p_3)$ is reached. If the consumer divides M between

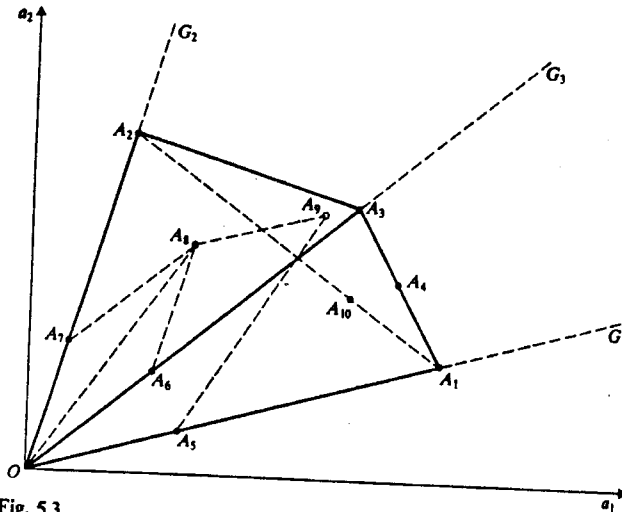


Fig. 5.3

goods 1 and 3 all points on the line A_1A_3 are attainable. For example, spending $M/2$ on each good produces the combination A_4 . Similarly, dividing M solely between goods 2 and 3 generates the combinations along A_2A_3 . Finally, spending M only on goods 1 and 2 produces combinations along A_1A_2 .

What happens if the consumer buys some of all three goods? Suppose she spends $M/3$ on each of the goods. $M/3$ spent on x_1 produces A_5 , on x_3 produces A_6 and on x_2 produces A_7 . The sum of A_7 and A_6 is A_8 , and of A_8 and A_5 is A_9 . A_9 is, however, an inefficient combination since by rearranging expenditure it is possible to increase the amount of at least one characteristic without reducing the other. For example, spending *all* of M on good 3 produces A_3 , which contains more of both characteristics than A_9 . This is not a result of the specific proportions of M spent on the three goods. Any bundle of goods containing all three goods will be inefficient: it will be possible to increase the amount of one characteristic without reducing the other. When all three goods are bought a point *below* the line $A_1A_3A_2$ will be produced. Hence efficient bundles of goods contain at most two goods. This result can be generalized: *if there are r characteristics and n goods and $n > r$, efficient bundles of goods will contain at most r goods.*

In our two characteristics case at most two goods will be bought, but which two? In the example shown in Fig. 5.3 the answer is either goods 1 and 3 or goods 2 and 3. Goods 1 and 2 can only produce combinations along A_1A_2 and hence for any bundle containing goods 1 and 2 it is possible to find a bundle containing goods 1 and 3 or 2 and 3 which produces more of both characteristics. For example, the bundle of goods 1 and 2 producing A_{10} is inefficient because by spending $M/2$ each on goods 1 and 3 A_4 , containing more of a_1 and a_2 , can be reached. Hence the efficient bundles of goods are those (a) on which the consumer spends all her income and (b) contain either only goods 1 and 3 or only 2 and 3.

Suppose that the price of good 3 is increased. The effect will be to shift the upper boundary of the consumer's feasible set inwards in both goods and characteristics space:

fewer bundles of goods and combinations of characteristics can be purchased. Figure 5.4 shows the effect of rises in the price of good 3 on the set of attainable characteristics combinations. The upper boundary is shifted inwards from $A_1A_3A_2$ to $A_1A_4A_2$. If p_3 rises sufficiently far the attainable point on OG_3 will lie below the line A_1A_2 . In this case bundles containing goods 1 and 3 will produce combinations along $A_1A_5A_2$. Such bundles will produce less of both characteristics than bundles containing only goods 1 and 2: $A_1A_5A_2$ lies below A_1A_2 . Hence, if the price of good 3 rises to this extent none of good 3 will be bought because that is inefficient.

Introduction of a new good

Figure 5.4 can also be used to examine the effect of the introduction of a new good. Assume there are initially only goods 1 and 2 and that good 3 is now introduced onto the market. This may extend the consumer's feasible set, pushing the upper boundary out from A_1A_2 to $A_1A_3A_2$. Previously some consumers bought goods 1 and 2, now some buy 1 and 3 and attain points along A_1A_3 , others buy goods 2 and 3 and attain combinations along A_2A_3 . If good 3 gives rise to a line like $A_1A_5A_2$ then it will be quickly withdrawn from the market since no consumer will buy it.

We can use our earlier definition of the implicit prices of characteristics to state the condition under which a new good will certainly not succeed in the market. The important point here is that this can be done independently of consumer preferences and indeed of income also. As Fig. 5.4 suggests, a new good will certainly fail if its prices and attribute bundle are such that they imply a point below the existing efficiency frontier. We can show that this is equivalent to the following. Suppose that we have the implicit prices π_1 and

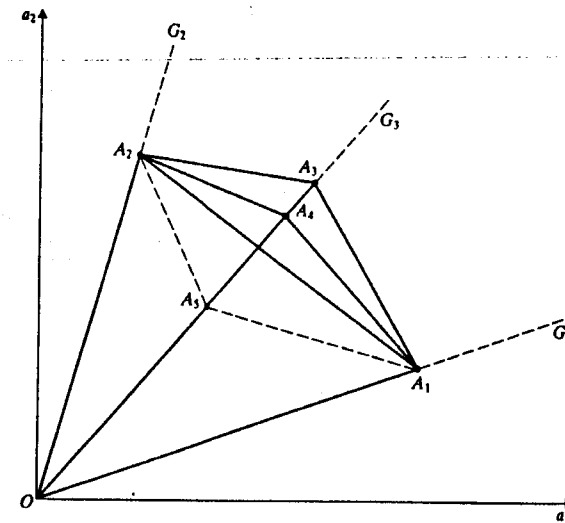


Fig. 5.4

π_2 found as in [B.8] or [B.10]. Then if

$$\pi_1 \alpha_{13} + \pi_2 \alpha_{23} < p_3 \quad [\text{B.11}]$$

the new product will fail. That is, if the value of its characteristics bundle at the existing implicit prices is less than its price per unit, then it yields a point such as A_5 in Fig. 5.4. The simplest way to show this is to multiply through [B.11] by $M/p_3\pi_2$ to obtain

$$\frac{M\alpha_{13}}{p_3} \frac{\pi_1}{\pi_2} + \frac{M\alpha_{23}}{p_3} < \frac{M}{\pi_2} \quad [\text{B.12}]$$

From [B.8], the equation of the line A_1A_2 , we have $M/\pi_2 = a_2 + a_1(\pi_1/\pi_2)$ and so [B.12] becomes (using also $a_{i3} = M\alpha_{i3}/p_3$, $i = 1, 2$)

$$(a_{13} - a_1)(\pi_1/\pi_2) + (a_{23} - a_2) < 0$$

Multiplying this through by π_2 and rearranging gives

$$\pi_1 a_{13} + \pi_2 a_{23} < \pi_1 a_1 + \pi_2 a_2 = M \quad [\text{B.13}]$$

implying that the point $A_5 = (a_{13}, a_{23})$ must lie below the original efficiency frontier A_1A_2 .

As well as having a nice economic interpretation, we could imagine the condition in [B.11] actually being empirically applicable. For example, if a firm can compute the existing implicit prices π_1, π_2 from market data, it can obtain the *maximum* price $\hat{p}_3 = \pi_1 \alpha_{13} + \pi_2 \alpha_{23}$ at which a new good would stand a chance in the market. If \hat{p}_3 is less than the unit cost of producing the new good it is pointless to go ahead. In the converse case further market research is worthwhile. Note that we cannot conclude that a good will *necessarily* succeed in the market if its cost is less than \hat{p}_3 . Whether consumers will actually buy the good depends on their preferences: the fact that a good is on the efficient frontier does not imply it will be bought, although if a good is not on the frontier it certainly will not be bought. We now need to close the model therefore by introducing preferences. Before doing so, we note also that the relevant implicit prices will vary with the particular linear segment of the efficiency frontier considered. For example, in Fig. 5.4, if the frontier is $A_1A_3A_2$, then implicit prices along A_1A_3 are found by solving

$$\pi_1 \alpha_{11} + \pi_2 \alpha_{21} = p_1$$

$$\pi_1 \alpha_{13} + \pi_2 \alpha_{23} = p_3$$

and those along A_3A_2 by solving

$$\hat{\pi}_1 \alpha_{13} + \hat{\pi}_2 \alpha_{23} = p_3$$

$$\hat{\pi}_1 \alpha_{12} + \hat{\pi}_2 \alpha_{22} = p_2$$

and in general $\pi_i \neq \hat{\pi}_i$, $i = 1, 2$. (See Question 8, Exercise 5B.)

Choice of the optimal bundle

The first stage of the consumer's problem was finding all efficient bundles of goods which generate the upper boundary of the feasible set in characteristics space. No information on preferences is required for this part of the problem, but solving it does not yield a

unique bundle of goods (except when the efficient set of bundles consists of a single bundle). The second part of the problem is choosing the optimal bundle from the set of efficient bundles and for this we require information on the consumer's preferences. Since we have assumed that the consumer's preferences over characteristics satisfy the assumptions of section 3A relating to preferences over goods, we can analyse the consumer's choice by superimposing her indifference map on Fig. 5.5. Each indifference curve shows the combinations of characteristics which are equally valued by the consumer. Since she is assumed to prefer more of both characteristics to less (they have positive marginal utility), the indifference curves are negatively sloped, and higher curves denote preferred combinations.

The optimal combination of characteristics is A^* where the highest indifference curve I_1 in the feasible set OA_1A_2 is reached. The optimal bundle of goods which produces A^* is found by drawing lines parallel to OG_1, OG_2 back from A^* to OG_2 and OG_1 at A_3, A_4 . The amount of x_1 in the optimal bundle is then proportional to the distance OA_3 and x_2 is proportional to OA_4 .

In this particular solution I_1 is tangent to A_1A_2 . We have shown that the slope of A_1A_2 can be interpreted as the negative of the ratio of the implicit prices π_1, π_2 of the characteristics. The consumer's preferences can be represented by a utility function $u(a_1, \dots, a_r)$ and so the slope of the indifference curve is

$$\left. \frac{da_2}{da_1} \right|_{du=0} = -\frac{u_1}{u_2}$$

where u_i is marginal utility of characteristic i . MRS_{21}^u is the marginal rate of substitution between characteristics 1 and 2: the rate at which the consumer is prepared to substitute

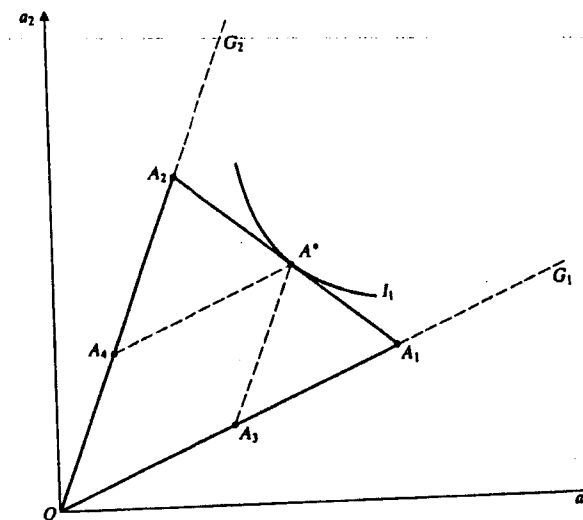


Fig. 5.5

characteristic 2 for characteristic 1. Hence the tangency solution to the consumer's problem satisfies

$$MRS_{21}^a = \frac{u_1}{u_2} = \frac{\pi_1}{\pi_2} \quad [B.14]$$

As in Chapter 3 there may also be a corner solution where [B.14] does not hold. For example, the highest attainable indifference curve may touch the feasible set as at A_1 . In this case only good 1 is bought.

The reader should superimpose the indifference map on the three-good case illustrated in Fig. 5.4 and examine the different types of solution. With a tangency solution (along A_1A_3 or A_2A_3) two goods are bought, but with a corner solution (at A_1 , A_2 or A_3) only one good is bought. Notice that some consumers may choose bundles with goods 1 and 3 and some bundles with goods 2 and 3 depending on their indifference map but none, irrespective of their preferences, will buy goods 1 and 2.

Comparative statics

As in the conventional model of Chapter 3 the bundle of goods chosen by the consumer will depend on her preferences, which determine her indifference map, and on money income M and the prices of the goods, which determines the set of feasible bundles. But, unlike the conventional model, choice is also affected by the consumption technology which determines the combinations of characteristics producible from a given bundle of goods. The consumer's demand functions are therefore of the form

$$x_i = D_i(p_1, \dots, p_n, M, a_1, \dots, a_n) \quad (i = 1, \dots, n)$$

The effect of changes in prices, income and technology on the demand for characteristics and hence goods can be established by making the appropriate changes in the feasible set and examining the consumer's possible responses with different types of preferences (indifference maps). The results are similar to those of Chapter 4 and are left as an exercise for the reader.

Exercise 5B

- 1.* The assumptions about $u(x)$ in section 3A ensure that the demand functions for x in section 3D are continuous and single valued (only one bundle maximizes u at each p, M combination). The $u(a)$ function used in this section also reflects the assumptions of section 3A. Does this imply that the demand functions for goods derived from the analysis here are continuous? Draw diagrams to illustrate the consumer's response to income and price changes in which the demand for good 1 varies (a) continuously and (b) discontinuously with M and p_1 .
2. Under what circumstances will the introduction of a new good raise, lower or leave unchanged the demand for an existing good?
3. The degree of substitutability between two goods could be defined, without reference to preferences, in terms of the angle between the rays in characteristics space

generated by the two goods. The smaller the angle the closer the substitutes. How will the degree of substitutability affect the answer to the previous question?

4. Show how the introduction of a new good may drive some existing goods off the market. Since the consumer now has fewer goods to choose from is she worse off?
- 5.* Discuss critically the assumption that the bundle of quantities of characteristics represented by one unit of a good is capable of objective measurement and is the same for all consumers. Illustrate your answer with examples of real consumer goods.
- 6.* Show that the following comparative static results can be established by simple diagrammatic analysis:
 - (a) the demand for characteristics and for goods may rise, fall or remain constant in response to changes in prices and income;
 - (b) the elasticity restrictions on the demand for goods of Question 4, Exercise 3D continue to hold;
 - (c) the demand for goods and for characteristics is homogeneous of degree zero in prices and income;
 - (d) demand for all characteristics will increase in response to a sufficiently large increase in income;
 - (e) normality of all characteristics does not imply that all goods are also normal;
 - (f) some goods may be Giffen goods even though no characteristics are Giffen characteristics.
- 7.* Show how the consumption technology model can be used to examine the cost imposed on the consumer by misleading advertising which causes the consumer to overestimate the output of characteristics from a good.
8. There are three goods currently sold in the market, with characteristics vectors (3,1), (2,2) and (1,3) respectively. Their prices are respectively 5, 4, and 5. Sketch the market efficiency frontier. A firm is considering entering the market with a new product whose characteristics vector is (2.5,1.5). This good will cost 4.50 per unit to produce. Is it worth attempting entry?

C. The consumer as a labour supplier

Our analysis in this and the two previous chapters has been concerned with the consumer's allocation of income among goods and has ignored the question of how the consumer allocates the time available in a given period. The problem is important. First one of the main sources of the income spent on the goods consumed is the sale of the consumer's time in return for a wage. Second, time is a scarce resource and the consumption of goods requires an input of time as well as of money. In this section we will examine a simple model in which the consumer chooses the amount of time spent at work. In the following section we will enquire more closely into how the time not spent at work ('leisure' time)

is allocated to the consumption of different goods and how this affects the consumer's labour supply decision.

The consumer's utility function is assumed to depend on the bundle of goods consumed (x) and the amount of non-work time or leisure (L).

$$u = u(x, L) \quad [C.1]$$

Since more leisure is assumed to be preferred to less, the marginal utility of leisure u_L is positive. The consumer is constrained in two ways. First she cannot spend more than her income M

$$\sum p_i x_i \leq M = wz + \bar{M} \quad [C.2]$$

where z is the length of time spent at work, w is the wage rate (assumed constant) and \bar{M} is non-work income from shares in firms, bond interest, government subsidies, etc. Since the marginal utility of every good is always positive (non-satiation assumption), [C.2] will be treated as an equality.

Second, the consumer in any given period of length T is constrained by her 'time budget'

$$T = z + L \quad [C.3]$$

which says that the time she has available is divided between work and leisure. The consumer's problem is to maximize $u(x, L)$ subject to [C.2] and [C.3] by choice of x , L and z .

One way to proceed would be to use [C.3] to substitute $T - L$ for z in the constraint on expenditure [C.2] and to rearrange [C.2] as

$$\sum p_i x_i + wL \leq \bar{M} + wT \equiv F \quad [C.4]$$

F is the individual's *full income*: the amount she would be able to spend if she used all her time endowment T to earn income. The problem of maximizing $u(x, L)$ subject to [C.4] is formally identical to the consumer problem studied in earlier chapters. We will therefore relegate this approach to the exercises and merely note that when the problem is set up in this way the wage rate w is clearly seen to be a price attached to the consumer's consumption of leisure.

In order to concentrate on the supply of labour we adopt instead a two-stage approach to the problem. At the first stage, the consumer's income M and her allocation of time between leisure and work are held constant. The consumer chooses her consumption bundle x to maximize $u(x, L)$ subject to the expenditure budget constraint [C.2] with L and M fixed. The optimal x at this stage depends on the prices of goods and on the fixed levels of L and M : $x^* = x(p, M, L)$. Substituting the optimal x^* into the utility function gives the indirect utility function:

$$\theta = u(x^*(p, M, L), L) = \theta(p, M, L) \quad [C.5]$$

which shows the maximum utility the consumer can achieve given the prices of consumption goods and the fixed levels of M and L . For given L , $\theta(p, M, L)$ has all the properties of the indirect utility functions discussed in previous chapters. Thus θ_M can be interpreted as the marginal utility of expenditure or income. The reader should write down the Lagrangean

for the first-stage problem and use the Envelope Theorem of section 2J to check that

$$\theta_L(p, M, L) = u_L(x^*, L) \quad [C.6]$$

At the second stage of the problem the consumer chooses her allocation of time to maximize the first stage indirect utility function. To emphasize the labour supply decision we use the time constraint [C.3] to substitute $T - z$ for L in [C.5]. Since we assume that the prices of consumption goods are constant we can write the consumer's objective function for her second-stage problem as

$$v = v(M, z) \equiv \theta(p, M, T - z) \quad [C.7]$$

from which it is clear that the marginal utility of work time is just the negative of the marginal utility of leisure time:

$$v_z(M, z) = -\theta_L(p, M, T - z) \quad [C.8]$$

The second-stage problem is to choose M and z to maximize $v(M, z)$ subject to $M = \bar{M} + wz$. Rather than set this problem up as a Lagrangean we substitute the constraint into the objective function, thereby transforming the two-choice variable-one-constraint problem into the unconstrained single-choice variable problem.

$$\max_z v(\bar{M} + wz, z) \equiv f(z; \bar{M}, w) \quad [C.9]$$

We assume that the consumer will always supply some labour and never spend all her time working so that we can ignore the direct constraints $z \geq 0$ and $z \leq T$. The first-order condition on z is

$$f_z(z; \bar{M}, w) = v_M w + v_z = 0 \quad [C.10]$$

Our two-stage approach is equivalent to the one-stage problem of maximizing the strictly quasi-concave direct utility function $u(x, L)$ subject to a linear budget constraint. Since the first-order conditions for the one-stage problem are necessary and sufficient to identify a unique global solution we know that [C.10] is also necessary and sufficient and that the second-order condition $f_{zz}(z; \bar{M}, w) < 0$ holds when [C.10] holds.

Figure 5.6(a) shows the consumer's feasible set in terms of z and M . The upper boundary of the feasible set is the *wage line* $\bar{M}M_1$, which plots income as a function of labour supplied $M = \bar{M} + w_1 z$, given the wage rate w_1 . The slope of the wage line is

$$\frac{dM}{dz} = w_1$$

so that increases in w_1 steepen the wage line. Increases in unearned income \bar{M} shift the intercept of the wage line upward but do not alter its slope. The indifference curves I_1 , I_2 and I_3 reflect our assumption that the consumer's utility function is strictly quasi-concave and since $v_M > 0$, $v_z < 0$ the indifference curves are positively sloped and higher indifference curves correspond to higher utility levels. The slope of the consumer's indifference curve is

$$\left. \frac{dM}{dz} \right|_{dv=0} = -\frac{v_z}{v_M} \quad [C.11]$$

The indifference curves reflect the physiological fact that the consumer cannot survive without some leisure (for sleeping, eating and so on) since they never intersect the line $z = T$ but get steeper as z increases.

When $w = w_1$ the solution to the labour supply problem is at A where the indifference curve I_1 is tangent to the wage line:

$$\left. \frac{dM}{dz} \right|_{dv=0} = -\frac{v_z}{v_M} = \frac{v_L}{v_M} = w_1 \quad [\text{C.12}]$$

The amount of money the consumer is *willing* to accept for a unit reduction in leisure (increase in labour supplied) is her marginal rate of substitution between income and leisure and this is equated to the amount of money she will *actually* receive for a unit reduction in leisure (increase in work time). This intuition is confirmed by our earlier formal analysis since rearrangement of the first order condition [C.10] yields [C.12].

Comparative statics: the labour supply curve

From [C.10] or from the figure it is clear that the optimal labour supply depends on the individual's unearned income and the wage rate, so that the *labour supply function* is

$$z^* = z(\bar{M}, w) \quad [\text{C.13}]$$

As the wage increases in Fig. 5.6 the wage line pivots about \bar{M} and becomes steeper and the optimal position changes from A to B and then to C as the wage increases from w_1 to w_2 and then to w_3 . In part (b) of Fig. 5.6 the labour supply curve shows the amount of labour supplied at the different wage rates, with points a , b and c on the supply curve S corresponding to the optimal positions A , B and C in part (a). The locus of optimal points in part (a) generates the supply curve in part (b). (Compare the relationship between the price consumption curve and the demand curve in Fig. 15 in Chapter 3.)

In Fig. 5.6 there is a *backward bending supply curve* with increases in w increasing the supply of labour at low wage rates but decreasing it at high wage rates. Since decreases in labour supplied imply increases in leisure demanded and w is the price of leisure, this apparently perverse response at high wage rates is analogous to a Giffen consumption good where an increase in price leads to an increase in demand. As with the Giffen good it is helpful to examine the effect of the change in wage rate in more detail.

The effect of changes in w on z supplied can be decomposed into income and substitution effects, as in the earlier analysis of the effects of changes in p_i on x_i demanded in section 3D. The 'wage effect' is the movement from A to B in Fig. 5.7. The wage line is then shifted downward parallel with itself until it is tangent to the initial indifference curve I_1 at D . The substitution effect AD shows the change due solely to the variation in w with utility held constant. DB is the income effect, showing the change due to the rise in utility with w held constant. $\Delta \bar{M}$ is the compensating variation in unearned income which will leave the consumer just as well off after the wage rise as she was before with her initial unearned income \bar{M} . The substitution effect of a wage rise is always to *increase* the supply of labour.

The wage line becomes steeper, and since the slope of I_1 rises as z rises, the point of tangency D between the wage line with slope w_2 and I_1 must be to the right of A .

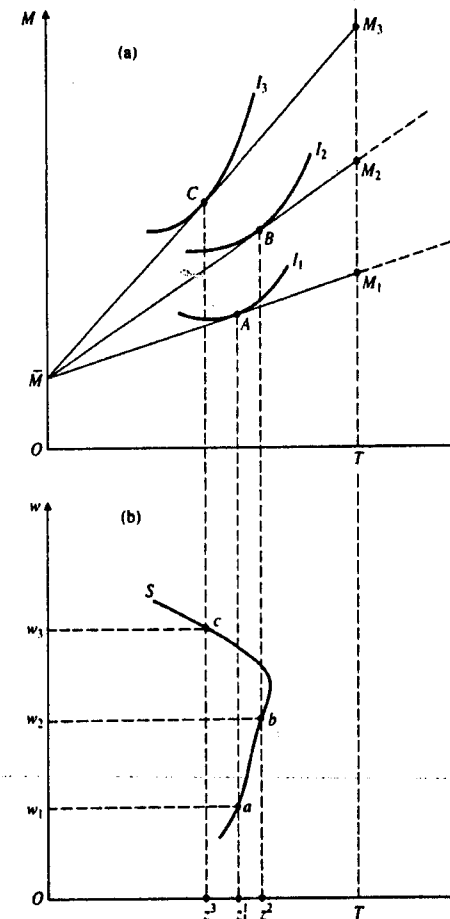


Fig. 5.6

No such restriction can be placed on the income effect. B may be to the right or (as in the figure) to the left of D . If B is to the left of D then z falls as income rises with constant w , or equivalently L rises as income rises, so that *leisure is a normal good*. Notice that if the supply of labour declines as w rises B in Fig. 5.7 must be to the left of A . Since A is always to the left of D the supply of labour would then *have* to fall (L rise) as income rises with constant w . Hence leisure being a normal good is a necessary, but not sufficient, condition for a backward bending, negatively sloped supply curve of labour.

That leisure is a normal good is plausible. Further, changes in the wage rate have a larger effect on the consumer's real income or utility than changes in the price of a consumption good, since expenditure on a consumption good is typically a small proportion of the consumer's income, whereas her earned income is likely to be a large proportion of her

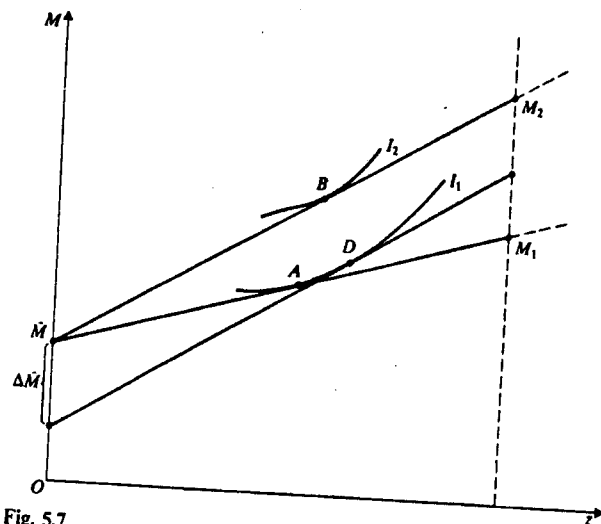


Fig. 5.7

total income. Hence a backward bending supply curve for labour seems to be more likely than an upward sloping demand curve for a consumption good.

The Slutsky equation. We can use the simple comparative static method outlined in section 2I to reinforce the conclusions of the diagrammatic analysis. To illustrate the power of this method we will use it to derive the Slutsky equation for labour supply. Our derivation is in three steps, with the first two steps also providing information on other aspects of the consumer's labour supply problem. First, we examine the effect of changes in unearned income on labour supply. Then we consider how unearned income must be changed as w varies so as to keep the consumer's utility constant. This enables us in the last step to distinguish the substitution and income effects of the change in the wage rate.

(a) The first-order condition [C.10] is an implicit function in the endogenous variable z and the exogenous variables \bar{M} and w . Changes in \bar{M} or w lead to changes in z so that [C.10] continues to hold. Thus holding w constant and totally differentiating $f_z(z; \bar{M}, w)$ with respect to \bar{M} and z gives

$$df_z(z; \bar{M}, w) = f_{zz}dz + f_{zM}d\bar{M} = 0$$

and rearranging we have

$$\frac{dz}{d\bar{M}} = -\frac{f_{zM}}{f_{zz}} \quad [C.14]$$

which is the effect on labour supply of an increase in unearned income with the wage rate held constant. The second-order condition $f_{zz} < 0$ implies that the sign of [C.14] is determined entirely by the sign of

$$f_{zM} = v_{MM}w + v_{zM} \quad [C.15]$$

It is apparent from Fig. 5.7 that after an upward parallel shift in the wage line from \bar{M}_1 the consumer will increase her labour supply if and only if her indifference curves become flatter as M increases with z constant, i.e. along vertical lines in M, z space. The reader should draw some diagrams to check this and also to confirm that if the indifference curves become steeper along such lines then the supply of labour is reduced. The partial derivative of [C.11] with respect to M shows the effect on the slope of the indifference curve of an increase in M with z held constant:

$$\begin{aligned} \frac{\partial[-v_z(M, z)/v_M(M, z)]}{\partial M} &= \frac{-[v_M v_{zM} - v_z v_{MM}]}{(v_M)^2} = \frac{-[v_{zM} - (v_z/v_M)v_{MM}]}{v_M} \\ &= \frac{-[v_{zM} + wv_{MM}]}{v_M} = \frac{-f_{zM}}{v_M} \end{aligned} \quad [C.16]$$

where the penultimate step uses [C.12]. Thus, since we cannot assume that the indifference curves become flatter or steeper along vertical lines in M, z space, we cannot determine the sign of f_{zM} and thus the effect of increases in \bar{M} on labour supply. Introspection suggests that leisure is a normal good: an increase in unearned income leads to an increased demand for leisure. Since increases in z correspond to reductions in L and vice versa, the plausible supposition that leisure is a normal good suggests that increases in unearned income will reduce labour supply. Note that an increase in \bar{M} leads to shifts in the labour supply curve rather than movements along it.

(b) The maximized utility attained after an optimal choice of labour supply is shown by the indirect utility function

$$v^* = v(\bar{M} + wz^*, z^*) = v(\bar{M} + wz(\bar{M}, w), z(\bar{M}, w)) = v^*(\bar{M}, w) \quad [C.17]$$

Increases in \bar{M} obviously make the consumer better off: $v_M^* = v_M > 0$. So do increases in w : differentiating v^* with respect to w gives a form of Roy's identity

$$v_w^* = v_M(M + wz^*, z^*)z^* = v_M^*z^* > 0 \quad [C.18]$$

Recall from sections 3D and 4B that the substitution effect of a price change is the change in demand induced by the change in relative prices with utility held constant by a compensating change in income. Consider what compensating variation in unearned income is necessary to keep the consumer's utility constant when w varies. Setting the total differential of v^* equal to zero

$$dv^* = v_M^* d\bar{M} + v_w^* dw = 0$$

Using Roy's identity for v_w^* and rearranging gives the rate at which \bar{M} must be varied to keep the consumer's utility constant when the wage rate increases:

$$\left. \frac{d\bar{M}}{dw} \right|_{dv^*=0} = -z^* \quad [C.19]$$

Alternatively, we can interpret z^* as the rate at which the consumer's real income increases when the wage rate increases.

(c) Applying the simple comparative static technique to the effects of an increase in w with \bar{M} held constant gives

$$\frac{dz}{dw} = \frac{-f_{zw}}{f_{zz}} \quad [C.20]$$

Now referring to [C.10], [C.15] and [C.19] we see that

$$f_{zw} = \frac{\partial [v_M(\bar{M} + wz, z)w + v_z(\bar{M} + wz, z)]}{\partial w} = [v_{MM}w + v_{zM}] \frac{d\bar{M}}{dw} + v_M$$

$$= f_{zM}z^* + v_M \quad [\text{C.21}]$$

and substituting [C.21] into [C.20] gives the *Slutsky equation of labour supply*

$$\frac{\partial z^*}{\partial w} = \frac{-[v_M + z^*f_{zM}]}{f_{zz}} = \frac{-v_M}{f_{zz}} + \frac{z^*(-f_{zM})}{f_{zz}} = \frac{-v_M}{f_{zz}} + z^* \frac{\partial z^*}{\partial \bar{M}} \quad [\text{C.22}]$$

where we have used [C.14] to substitute $\partial z^*/\partial \bar{M}$ for $-f_{zM}/f_{zz}$. An increase in the wage rate has two effects on the supply of labour since it alters the relative prices of goods and leisure and it increases the consumer's real income or utility level. [C.22] decomposes the effect of a change in the wage rate on the supply of labour into two terms. The second of these is the income effect of the wage rate increase. The income effect is the product of two terms. The first, z^* , is a measure of the rate at which an increase in w increases the consumer's real income or utility. The larger is z^* the greater the effect of w on the consumer's utility and the greater the absolute magnitude of the income effect. The second, $\partial z^*/\partial \bar{M}$, measures how responsive the supply of labour is to an increase in unearned income. As we have seen this response may be positive or negative depending on whether leisure is an inferior or a normal good. The first term in [C.22] is the substitution effect: the change in labour supply induced by an increase in w with real income or utility held constant. As with the substitution effect in demand theory the substitution effect in labour supply is definitely signed since marginal utility of income v_M is positive and by the second-order condition $f_{zz} < 0$. Hence the substitution effect of a wage increase is always positive: the consumer will always increase her labour supply if the wage rate is increased and her utility held constant.

Exercise 5C

- What is the effect on
 - the feasible set and
 - the labour supply curve of
 - a proportional income tax; (ii) overtime payments; (iii) unemployment benefit; (iv) fixed hours of work?
- What is the effect on labour supply of replacing a proportional income tax with a progressive (increasing marginal rate) income tax which yields the same tax revenue?
- Target income.* Suppose that a worker has a target income: she supplies just enough labour to produce a particular total income. Sketch her indifference curves and her labour supply curve. What is the effect of an increase in unearned income?

- Slutsky equation and Hicksian labour supply function.* Consider the problem of minimizing the unearned income \bar{M} necessary to achieve a given utility level v .
 - Set up the Lagrangean and derive and interpret the first-order conditions.
 - Show that the derivative of the minimized unearned income $\bar{m}(w, v)$ is the constant utility (Hicksian) labour supply $\zeta(w, v)$.
 - Use $\bar{m}(w, v)$ to derive the Slutsky equation. (Use the fact that $\zeta(w, v) = z(m(w, v), w)$.)
- Economic rent.* Use the indirect utility function $v^*(\bar{M}, w)$ to define the compensating and equivalent variation measures of the benefit to the consumer of a change in the wage rate. What is the relationship between these measures and the consumer's Hicksian $\zeta(w, v)$ and Marshallian $z(\bar{M}, w)$ supply curves?
- Show that the two-stage approach problem in this section is equivalent to the one-stage problem of choosing x and L to maximize $u(x, L)$ subject to [C.4].
- Consumer prices and labour supply.* Assume that the consumer's utility function is weakly separable in goods and leisure. What effect will a change in the price of the i th consumption good have on her supply of labour?

D. Consumption and the allocation of time

In our discussions of the consumption decision so far we have assumed that the only requirement for the consumption and enjoyment of goods was money to purchase the goods. We now examine the implications of recognizing that the consumption of goods requires an input of the consumer's time, and that time is a scarce resource. Watching a film in a cinema, eating a meal or merely resting all require, in addition to the expenditure of money on cinema tickets, food or an armchair, an expenditure of time. Consumption decisions are therefore constrained by the time needed in the various consumption activities as well as by the consumer's money income. Increasing the time spent working will increase money income but will reduce the amount of time available for use in consuming the goods. The consumer's problem is therefore to allocate his time *and* his money income. We will consider a model which simultaneously examines the consumer's labour supply and consumption decisions.

The consumer's utility depends in the usual way on x , the bundle of goods consumed: $u = u(x)$. The consumption decision is constrained in two ways. First the bundle of goods consumed cannot cost more than the consumer's income: $\sum p_i x_i \leq M = \bar{M} + wz$. Second, there is a time constraint: $T = \sum T_i + z$, where T_i is time spent consuming good i and z is work time. For simplicity we assume that there is a proportional relationship between the amount of good i and the length of time used in its consumption

$$T_i = t_i x_i \quad [\text{D.1}]$$

where t_i is the 'time price' of good i : the number of minutes required for consumption of one unit of good i .

The consumer's problem is (ignoring the non-negativity constraints)

$$\max_x u(x) \text{ s.t. (i) } \sum p_i x_i \leq \bar{M} + wz$$

$$(ii) \sum t_i x_i + z = T \quad [\text{D.2}]$$

Notice that the T_i and z are not choice variables; this is because the proportionality assumption [D.1] implies that choice of a bundle of goods determines the length of time spent consuming each good and hence also determines $z = T - \sum t_i x_i$. It would be possible to relax the proportionality assumption for many goods and allow for time spent in consuming goods to enter the utility function directly. We could also use the Lancaster consumption technology model and assume that utility-yielding characteristics are produced by consumption activities which use time and goods as inputs. These developments would, however, make the model rather complex and so we will limit ourselves to examining the implications of our very simple assumptions.

Using the time constraint in [D.2] we have $z = T - \sum t_i x_i$ and substituting in the expenditure constraint gives

$$\sum p_i x_i \leq \bar{M} + w(T - \sum t_i x_i)$$

or

$$\sum p_i x_i + w \sum t_i x_i = \sum (p_i + wt_i) x_i = \sum p_i x_i \leq \bar{M} + wT = F \quad [\text{D.3}]$$

t_i is the time necessary for the consumption of one unit of good i and w is the money value of a unit of time so that wt_i is the *time cost*: the opportunity cost of the time used in consuming a unit of good i . p_i is the money price of the good so that $p_i = p_i + wt_i$ is the *full price* of good i . As in the previous section F is the *full income* of the consumer: the maximum potential income if all time is used for earning. [D.3] is the *full budget* constraint of the consumer: the full cost of the goods consumed cannot exceed the consumer's *full income*. We assume that the consumer is not satiated so that at least one good has positive marginal utility and the budget constraint will always bind at the solution. Hence we will treat [D.3] as an equality constraint.

In the two-good case the constraint is, from [D.3]:

$$(p_1 + wt_1)x_1 + (p_2 + wt_2)x_2 = \bar{M} + wT$$

or

$$x_2 = [\bar{M} + wT - (p_1 + wt_1)x_1] / (p_2 + wt_2)$$

and so the slope of the full budget line F is

$$\frac{dx_2}{dx_1} = -\frac{(p_1 + wt_1)}{(p_2 + wt_2)} \quad [\text{D.4}]$$

The full budget line is drawn as F in Fig. 5.8. We can also draw the money and time budget constraints in the figure. For example, B' shows all bundles costing $\bar{M} + wz'$ and B'' all bundles costing $\bar{M} + wz''$, where $z'' > z'$. Similarly all bundles along L' require a total time input in consumption of $T - z'$ and those along L'' a total consumption time input of $T - z''$. The B and L lines can be thought of as isoexpenditure and iso-leisure

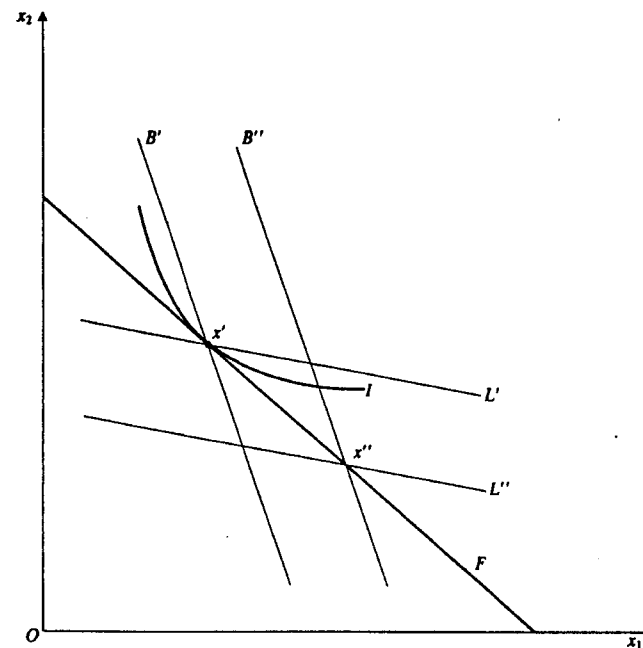


Fig. 5.8

contours. F is the locus of bundles satisfying both time and expenditure constraints simultaneously. For example, x' is on both B' and L' and x'' on both B'' and L'' .

As in Chapter 3, the slope of the isoexpenditure lines is $-p_1/p_2$. Since, for given z along the L lines, variations in x_1 and x_2 must satisfy $t_1 dx_1 + t_2 dx_2 = 0$, the slope of the iso-leisure lines is $-t_1/t_2$. Fig. 5.8 has been drawn so that

$$\frac{t_1}{t_2} < \frac{p_1}{p_2} \quad [\text{D.5}]$$

Good 1 is in this case relatively less expensive in terms of time than good 2 but relatively more expensive in terms of money. Alternatively, we can define the *time intensity* of the i th good as the proportion of the full price accounted for by the time cost; $wt_i/(p_i + wt_i)$. Writing [D.5] as $t_1 p_2 < t_2 p_1$, adding $t_1 wt_2$ to both sides and multiplying through by w yields

$$\frac{wt_1}{p_1 + wt_1} < \frac{wt_2}{p_2 + wt_2} \quad [\text{D.6}]$$

so that in Fig. 5.8 good 1 is less time-intensive than good 2. As the consumer moves down F he substitutes the less time-intensive x_1 for the more time-intensive x_2 . He thereby consumes bundles with a greater money cost but with a smaller leisure time input, leaving him more time to earn the extra income required to pay for the more costly bundles.

Equilibrium of the consumer

The consumer is assumed to have preferences which satisfy the assumptions of section 3A and so we can analyse his choice by superimposing his indifference map on his feasible set as in Fig. 5.8. In the tangency solution at x' shown here the slope of the indifference curve I is equal to the slope of the full budget line, or

$$\frac{u_1}{u_2} = \frac{p_1 + wt_1}{p_2 + wt_2} \quad [\text{D.7}]$$

The consumer's marginal rate of substitution is set equal to the ratio of full prices, rather than the ratio of money prices as in the model of Chapter 3. Choice of $x' = (x'_1, x'_2)$ determines total time spent on consumption ($t_1x'_1 + t_2x'_2$) and at work ($T - t_1x'_1 - t_2x'_2 = z'$) and the amount of income earned (wz'), which together with unearned income is just sufficient to buy the bundle chosen ($\bar{M} + wz' = p_1x'_1 + p_2x'_2$).

Comparative statics

The problem of maximizing the strictly quasi-concave $u(x)$ subject to $\sum p_i x_i = F$ is mathematically equivalent to the problem of maximizing $u(x)$ subject to $\sum p_i x_i = M$ which we examined in Chapter 3. We leave the reader to set up the Lagrangean and derive the first-order conditions. The Marshallian demands are a function of the vector of full prices ρ and the full income

$$x_i = x_i(\rho, F) \quad [\text{D.8}]$$

and so is the consumer's indirect utility:

$$v(\rho, F) = u(x(\rho, F)) \quad [\text{D.9}]$$

The reader should also check that the problem of minimizing the full cost $\sum p_i x_i$ of achieving a given utility level u will yield the Hicksian demands $h_i(\rho, u)$ and the full cost or expenditure function $c(\rho, u) = \sum p_i h_i(\rho, u)$.

We also leave it to the reader to apply the techniques of Chapter 3 to investigate the effects of changes in ρ and F (just replace p and M with ρ and F in the steps in the arguments). Instead we examine the implications of a change in the wage rate. Since w influences all the full prices ($\rho_i = p_i + wt_i$) and the full income ($F = \bar{M} + wT$) its effects are more complicated than a change in income or a single price. Consider first the effect of w on the consumer's maximized utility. Allowing for its effects on the full prices and full income the partial derivative of the indirect utility function [D.9] with respect to w is

$$\frac{\partial v(\rho, F)}{\partial w} = \sum \frac{\partial v}{\partial \rho_i} \frac{\partial \rho_i}{\partial w} + \frac{\partial v}{\partial F} \frac{\partial F}{\partial w} = -\sum v_F x_i t_i + v_F T = v_F [T - \sum t_i x_i] = v_F z \quad [\text{D.10}]$$

where we have used Roy's identity for the effects of the full prices on v : $v_i = \partial v(\rho, F) / \partial \rho_i = -v_F x_i(\rho, F)$. [D.10] is just another version of Roy's identity: the effect on utility of an increase in the price of a commodity (labour) that the individual sells is the quantity sold ($z = T - \sum t_i x_i$), which is the rate at which income increases with price, times the marginal utility from additional income.

The effect of w on the consumer's demand for goods is also complicated because w alters all full prices and full incomes. Recall from section 4B that the Marshallian and Hicksian demands are equal if the required utility level in the full cost minimization problem is set at the maximized utility achieved in the utility maximization problem:

$$x_i(\rho, F) = h_i(\rho, v(\rho, F)) \quad [\text{D.11}]$$

Hence the effect of full income on the Marshallian demand for good i is

$$x_{iF}(\rho, F) = h_{iv} v_F \quad [\text{D.12}]$$

and the effect of w on x_i is

$$\frac{\partial x_i(\rho, F)}{\partial w} = \sum_j \frac{\partial h_i}{\partial \rho_j} \frac{\partial \rho_j}{\partial w} + h_{iv} \frac{\partial v}{\partial w} = \sum_j h_{ij} t_j + h_{iv} v_F z = \sum_j h_{ij} t_j + z x_{iF} \quad [\text{D.13}]$$

where $h_{ij} = \partial h_i / \partial \rho_j$ is the cross substitution effect of the full price ρ_j on the constant utility (Hicksian) demand for good i . The effect of w on the demand for good i has been decomposed into an income effect ($z x_{iF}$) and the sum of n substitution effects. In general, the effect of an increase in the wage on the demand for goods is ambiguous for two reasons. First, the good may be normal or inferior so that the increase in real income or utility caused by the increase in the wage rate may increase or reduce demand. Second, the change in the wage rate alters *all* the relative full prices and so leads to n substitution effects, rather than just one as in the case of change in a single price.

If we assume that there are just two goods we can get some insight into the substitution effects of the wage increase. Consider the sum of the substitution effect terms in [D.13] in the case of good 1:

$$h_{11} t_1 + h_{12} t_2 \quad [\text{D.14}]$$

The Hicksian demand functions are homogeneous of degree zero (see section 4B) so that we can use Euler's Theorem (see section 7B) to establish

$$h_{11} \rho_1 + h_{12} \rho_2 = 0$$

This enables us to substitute $-h_{11} \rho_1 / \rho_2$ for h_{12} in [D.14] to get

$$h_{11} t_1 - h_{11} \frac{\rho_1}{\rho_2} t_2 = h_{11} \left[\frac{t_1}{\rho_1} - \frac{t_2}{\rho_2} \right] \rho_1 \quad [\text{D.15}]$$

Since the own full price substitution effect h_{11} is always negative, the wage substitution effect is also negative if good 1 has a greater time intensity than good 2 (recall our discussion of [D.6]). [D.15] indicates that in the two-good case the wage substitution effect on good 1 is proportional to the own full price substitution effect on good 1. The full price of good 2 falls relative to the full price of good 1 when w increases if good 1 is more time intensive than good 2 and so the wage substitution effect is in the same direction as the own full price substitution effect, leading to a decrease in demand for good 1. If good 1 is less time intensive than good 2 an increase in w is equivalent to a reduction in the relative full price of good 1 and so the wage substitution effect would increase the demand for good 1.

We can use [D.11], [D.12] and [D.13] to examine the effect of w on the consumer's supply of labour. Using [D.11], the consumer's Marshallian labour supply is

$$z(\rho, F) = T - \sum t_i x_i(\rho, F) = T - \sum t_i h_i(\rho, v(\rho, F)) = \zeta(\rho, v(\rho, F)) \quad [\text{D.16}]$$

where $\zeta(\rho, v)$ is the Hicksian constant utility labour supply function. An increase in full income changes labour supply at the rate

$$z_F(\rho, F) = -\sum t_i x_{iF}(\rho, F) = -\sum t_i h_{iv} v_F = \zeta_v v_F \quad [\text{D.17}]$$

which may be positive or negative.

The effect of w on labour supply is again more complicated because account has to be taken of the effects of changes in the full income and all the full prices on all the n demands for goods. Using [D.13] and [D.17]

$$\begin{aligned} \frac{\partial z(\rho, F)}{\partial w} &= -\sum_i t_i \frac{\partial x_i}{\partial w} = -\sum_i t_i \left[\sum_j h_{ij} t_j + z x_{iF} \right] = -\sum_i \sum_j t_i h_{ij} t_j - z \sum_i t_i x_{iF} \\ &= -\sum_i \sum_j t_i h_{ij} t_j + z z_F = \frac{\partial \zeta(\rho, v)}{\partial w} + z(\rho, F) z_F(\rho, F) \end{aligned} \quad [\text{D.18}]$$

which is the *Slutsky equation for labour supply*. The last term in [D.18] is the income effect whose sign we have already noted is ambiguous and whose magnitude depends on the amount of labour supplied. The first term is the own substitution effect of w on the supply of labour. The full cost function $c(\rho, u)$, like the expenditure function of section 4A, is strictly concave in full prices if the utility function is strictly quasi-concave and suitably smooth. Hence the quadratic form $\sum_i \sum_j t_i h_{ij} t_j$ is negative definite (see section 2I) and so the own wage substitution effect is positive.

Figure 5.9 illustrates the effects of an increase in w for the two-good case. A rise in w will affect the feasible set in two ways. First, it shifts the full budget line F outwards over at least some of its range, thus permitting the consumer to achieve a higher indifference curve. A rise in w will not affect the position of the iso-leisure lines but it will permit a larger expenditure ($\bar{M} + wz$) for a given z . Hence all the isoexpenditure lines will shift outward, with constant slope. At the consumer's initial equilibrium x^1 in Fig. 5.9 he will no longer be spending all his money income. He could therefore, without altering z , move down the initial iso-leisure line L^1 to its intersection with the money budget line \hat{B}^1 where, at the initial level of z the consumer is again just spending all his income. This point \hat{x}^1 is a point on the new full income budget constraint F^2 since it satisfies both the time constraint and the new money income budget constraint. The rise in w , as might be expected, makes the consumer better off, since \hat{x}^1 must lie above the initial indifference curve I^1 . (Explain why.)

The second effect of the rise in w is to change the slope of the full budget constraint. In Fig. 5.9 F^2 is flatter than the initial full income line F^1 . This is because the diagram is drawn to reflect the assumption that good 1 is less time intensive than good 2. Because good 1 has a lower time price relative to the money price than good 2 the opportunity cost of time used in consuming a unit of good 1 (wt_1) is a smaller proportion of the full price than is the case with good 2. Hence a rise in w will raise the full price of good 1 by proportionately less than for good 2.

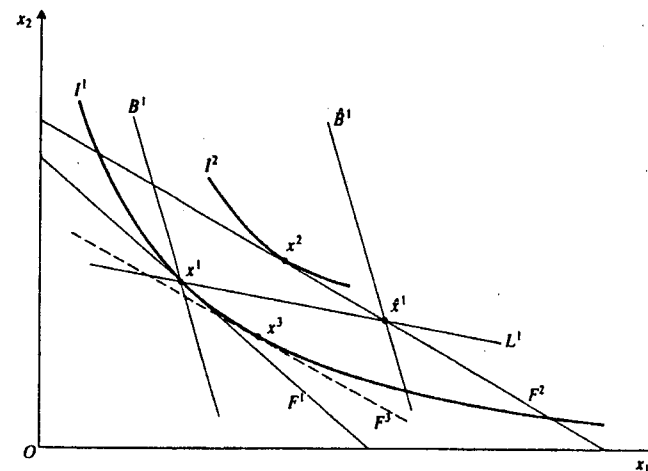


Fig. 5.9

The consumer's new optimum bundle on F^2 is x^2 , where more of both goods is consumed. The consumer's money income is larger at x^2 than at x^1 but x^2 lies on an iso-leisure line above L^1 indicating that more time is devoted to consumption and the consumer's supply of labour has been reduced by the rise in w . Only if the optimum was to the right of \hat{x}^1 on F^2 would the rise in w lead to an increased supply of labour as the consumer chose a less time-consuming consumption bundle.

As in Chapter 3, the comparative static responses to changes in the exogenous variables such as the wage rate will depend on the consumer's preferences. However, we can use Fig. 5.9 to decompose the effect of the change in w into an income effect and a substitution effect. As in previous analysis we move the new (full) budget line inward until the consumer can just achieve his initial level of utility on I^1 . Since the new budget line is flatter than the initial one the compensated demand for goods will be to the right of the initial equilibrium at x^3 where F^3 is tangent to the initial indifference curve. Hence we can establish that the substitution effect of a rise in the wage rate will be to increase the demand for the less time intensive good x_1 . This compensated change in w will lead to a rise in the amount of labour supplied (since x^3 is on a lower iso-leisure line than L^1). Thus the own substitution effect of the wage on the supply of labour leads to an increase in the supply of labour. Since the overall effect of the increase in w is to reduce the labour supply as the consumer moves from x^1 to x^2 the income effect is negative and sufficiently large to offset the substitution effect. Note that in the figure both goods are normal which explains the negative income effect of w on labour supply. (Is it necessary that both goods are normal for an increase in F to reduce labour supply?)

This analysis of the effect of a rise in the wage rate suggests some tentative explanations for two phenomena associated with rising real wages, i.e. with w rising faster than the money prices of goods. First, the substitution effect will lead to the substitution of goods which are less time-intensive for goods which are more time-intensive. Consumers will spend money in order to save time by buying higher-priced goods which have a smaller

time cost. Examples include 'convenience' foods which require less time for preparation and the greater use of domestic appliances to economize on time. Second, the secular decline in the average number of hours worked per worker may be ascribed to the strength of the income effect of rising real wages. This more than offsets the substitution effect and leads to an increase in leisure time used for consumption of the larger basket of goods bought with the rising full income.

Exercise 5D

1. Prove, as asserted in the text, that a rise in w will indeed flatten the full budget line if and only if good 1 is less time intensive than good 2. Investigate the effect on the demand for goods 1 and 2 and the supply of labour of an increase in the wage rate when good 1 is more time intensive than good 2. Show that the wage own substitution effect on labour supply is positive.
2. Sketch the effect on the feasible set of changes in unearned income, money prices and the length of time required for consumption. Examine the resulting changes in the optimal bundle chosen by the consumer.
3. Use the indirect utility function and the cost function to examine the implications for the consumer's utility and the consumer's behaviour of an increase in (a) p_i ; (b) t_i .
4. It is often suggested that individuals with larger incomes have smaller own money price elasticities of demand for goods. Give a rationale for this by examining the relationship between the money price elasticity of demand for a good and the consumer's wage rate.
5. How could the results of this section be used to evaluate the benefit to a consumer of, say, a new bridge which reduced the length of time he takes to get to work?

References and further reading

Revealed preference theory was developed in

P. A. Samuelson. 'A note on the pure theory of consumer behaviour', *Econometrica*, 5, 1938, 61–71.

Its usefulness for making aggregate welfare judgements was considered in

P. A. Samuelson. 'Evaluation of real national income', *Oxford Economic Papers*, 1950, 1–29.

Price indices are related to consumer theory in

A. Deaton and J. Muellbauer. *Economics and Consumer Behaviour*. Cambridge University Press, Cambridge, 1980, ch. 7,

which also has good discussions of the characteristics approach (ch. 10) and labour supply (ch. 11).

Consumption technology theory and its applications are set out in

K. Lancaster. *Consumer Demand. A New Approach*. Columbia University Press, New York, 1971.

The relationship between the consumer's preferences over characteristics and his demand for goods is discussed in

R. G. Lipsey and G. Rosenbluth. 'A contribution to the new theory of demand: a rehabilitation of the Giffen Good', *Canadian Journal of Economics*, 4, 1971, 131–163.

The seminal work on the economics of time is

G. S. Becker. 'Theory of the Allocation of Time', *Economic Journal*, 75, 1965, 493–517.

and the time and household production models are integrated in

R. T. Michael and G. S. Becker. 'On the new theory of consumer behaviour' *Swedish Journal of Economics*, 75, 1973, 378–396.

Both the preceding references are reprinted, along with other papers applying the same kind of models to a wide variety of behaviour, in

G. S. Becker. *The Economic Approach to Human Behaviour*. University of Chicago Press, Chicago, 1976.

CHAPTER 6

The firm

A. Introduction

In the last three chapters we have examined the theory of the consumer at some length. Our aim was partly to explain and predict consumer behaviour and partly to derive some general results, which will be used in Chapters 10 and 16 as building blocks in constructing a theory of markets and of resource allocation in the economy as a whole.

A *pure exchange economy* is one in which economic agents have given endowments of goods and exchange goods among themselves to achieve preferred consumption patterns. In analysing the determination of prices and quantities exchanged in such an economy, it would be sufficient to use the consumer theory so far constructed – the model of section 3F would be directly applicable. However, the pure exchange economy lacks an important aspect of real economies, namely *production*. Production is the activity of combining goods and services called *inputs*, in technological processes which result in other goods and services called *outputs*. In the pure exchange economy, although each consumer can transform his endowed bundle of goods into some other bundle through exchange, this is not true for the group as a whole: the sum of consumptions of each good cannot exceed the sum of initial endowments of it. The existence of production possibilities adds another dimension to economic activity: it permits transformations of endowed bundles of goods into other bundles for the economy as a whole. Clearly, any attempt at explaining resource allocation is incomplete unless it takes production into account. Theories of the firm arise out of the need to incorporate production into the theory of resource allocation. The firm is the institutional means by which production is organized in a market economy.

B. The nature of the firm

At a primitive stage in economic development production can be wholly individualistic, being carried on by one person working with tools and raw materials. Some goods and services are still produced in this way, for example, writing (though not publishing) a book, giving someone a haircut, painting a picture, but the overwhelming majority of

goods and services (including some books, haircuts and paintings) are produced by *co-operating groups* of individuals. The reasons for this are not hard to find: *specialization* of individuals in parts of the production process can be carried further within a group than if one individual undertakes the whole process and, in many processes, there are gains from *teamwork* – the total output of a group when working as a team is greater than the sum of outputs of individuals working separately. However, the 'firm' as it has been traditionally conceived of in economics is more than simply a co-operating group of producers; it is a group with a particular *organizational structure*, and a particular set of *property rights*. For example, it is possible to conceive of a 'producers' co-operative' in which assets are owned in common – no individual has the right to exclusive use or disposal of any of the equipment, output, cash reserves, etc. of the co-operative and decisions are taken by majority vote. This would clearly not be a 'firm' as traditionally conceived. The essence of the latter is the existence of a central figure, the owner, employer or *entrepreneur*, who:

- (a) enters into a contract with each of the individuals who supply productive services, which specifies the nature and duration of those services and the remuneration for them;
- (b) either takes decisions, or has the right to insist that decisions are taken, in *his* interests, subject to his contractual obligations;
- (c) has the right to the *residual income* from production, i.e. the excess of revenue over payments to suppliers of productive services made under the terms of their contracts;
- (d) can transfer his right in the residual income, and his rights and obligations under the contracts with suppliers of productive services, to another individual;
- (e) has the power to direct the activities of the suppliers of productive services, subject to the terms and conditions of their contracts;
- (f) can change the membership of the producing group not only by terminating contracts but also by entering into new contracts and adding to the group.

The essential feature of the 'classical firm' is therefore a central figure, with whom all contracts are concluded, and who controls and directs in his own interests, subject to constraints arising out of the terms of the contract he has made.

Since we can conceive of different ways of organizing co-operating productive groups, it is of interest to ask why this particular form, the entrepreneurial firm, developed into the dominant form of organization of production. It is possible to give an historical account of this: in the transition from the feudal, largely agrarian economy of the late Middle Ages to the capitalist industrial economy of the nineteenth and twentieth centuries, an important role was played by wealthy men who had accumulated their wealth through trade, inheritance of land, or by being successful skilled craftsmen. These were able to respond to important developments in transport and production technology, especially mechanization and the use of steam power, by investing in plant and machinery, grouping workers together into factories, entering into contracts of employment with them, and financing production in advance of sales. Thus their ownership of wealth was translated into their

ownership of the assets of the producing group and they became the buyers of labour power. The advent of the 'capitalist entrepreneur' thus shaped the organization of the producing group into that of the classical firm just described.

However, though this historical account may give a description of what happened, it does not constitute a complete explanation because it does not fully explain why it was this and not other forms of organization which came to dominate. Other organizational forms were certainly known and attempted, for example the early socialist experiments of Robert Owen. In addition, as R. Coase (1937) has emphasized, an alternative to organization within the firm which is always available is that of *organization through the market*. By this is meant the co-ordination of the myriad separate, individual decisions by the 'impersonal workings' of the price system. In this process there is no 'central planning' but only the self-interested planning of individual economic decision-takers, which interacts through the system of prices and markets to determine a resource allocation. In a phrase borrowed from D. H. Robertson, Coase describes the firm as 'an island of conscious power' in this 'ocean of unconscious co-operation'. Within the firm there is centralized economic planning and administrative co-ordination replaces the price mechanism, although, of course, the firm is embedded in an external system of market forces which condition its operations. The question then is: why does the firm, viewed as a centrally planned system, replace co-ordination through the market, and become the dominant form of organization of the producer group?

An explanation of the dominant position of the classical firm in the organization of production must rest on a demonstration of the advantages which it has over other forms of organization, including that of the market. Thus Coase has argued that the firm superseded market organization because there are costs associated with use of the price mechanism and that administrative organization within the firm is, up to a point, less costly. The major types of costs involved in effecting market transactions are those of acquiring information about prices and terms under which trade takes place; the costs of negotiating, writing and enforcing contracts; and the uncertainty which may exist about the conditions on markets in the future. In some kinds of activities and markets these costs might be minor, but in others they could greatly exceed the costs of organizing production within a firm, in which case we would expect the latter to dominate.

A second important reason for the dominance of the classical firm, not only over organization through the market but also over other forms of organization, such as that of the 'producer co-operative', has been advanced by Alchian and Demsetz (1972). When the producing group works as a team, there is the problem of measuring and rewarding each member's effort in such a way as to reward high productivity and penalize 'shirking'. In the absence of such measurement and reward, the presumption is that it pays any one individual to minimize his effort, since the costs of doing so, in terms of reduced output, are spread over all the members of the team – this is an example of the 'free rider problem' of Chapter 18. Then, it is argued, the system under which a central individual monitors performance and apportions rewards stimulates productivity, as the retention by that individual of the residual income of the group provides an incentive for him to perform the monitoring function efficiently. To this we might add that in terms of the speed with which decisions are arrived at, the costs absorbed in the decision-taking process, and the flexibility of response to changed circumstances, a system based on central direction rather than multilateral consultation and voting procedures is likely to have an advantage. (See

the discussion in Chapter 18 on the problems of common access resources and voting procedures, and relate it to the question of the likely efficiency of a 'producers' co-operative'.) It can therefore be argued that the classical 'entrepreneurial firm' emerged as the dominant form of organization of production because it had advantages of efficiency and productivity over other forms of organization, whether the market or formal organizations with different systems of decision structure and property rights.

The characteristics of the 'classical firm' described above have determined the form of the 'theory of the firm' in economics, and hence the representation of how production is carried on in a private ownership economy. The firm is viewed as being faced with an optimization problem (see Chapter 2). Its choice variables are input and output levels and possibly other variables such as advertising and expenditure on research and development. Its objective is to maximize profit, defined as the excess of revenue over all opportunity costs, including those associated with the supply of capital and the managerial functions of planning, organizing and decision-taking. This formulation of the objective function appear quite natural, since the individual controlling the firm receives the profit as his income (over and above payment for his supply of productive services), and, viewing him as a consumer, his utility from consumption is greater the greater the income available for it.

The constraints in the problem are of two types. First, the conditions on the markets which the firm enters as a seller of outputs or buyer of inputs will determine, through prices, the profitability of any production plan (i.e. a particular set of quantities of inputs and outputs), and hence also the way in which profits vary with the production plan. Another way of putting this is to say that market conditions determine the terms of the contracts into which the firm enters with buyers and suppliers of goods and productive services, and hence determine the amount of the residual income, or profit, which can be made. Second, the state of technology will determine which production plans are feasible, i.e. what amounts of inputs are required to produce given output levels, or conversely what outputs can be produced with given input levels. Thus, market conditions on the one hand, and the nature of technology on the other, determine the constraints in the firm's optimizing problem.

The classical theory of the firm operates at a high level of abstraction, at least equivalent to that of the theory of the consumer in Chapter 3. In its basic formulation, the firm is assumed to know with certainty the market conditions and state of technology. The theoretical problem is then to formalize the firm's optimization problem; examine the nature of its solution and the way in which this solution varies with changes in the parameters of the problem; and then translate the results into explanations and predictions of the firm's behaviour. Before going on to examine all this in detail, we consider in the next two sections some criticisms and extensions of this approach.

Exercise 6B

1. Employment in the UK ports industry used to be subject to the 'casual system'. Twice each day, dock-workers and employers would assemble at a particular place at each port, and employers would hire the men they wanted for a specific job, the men being paid off once the job was completed, possibly the same day or a little later. Explain why this system of allocating labour resources in the ports industry could be called 'co-ordination by the market'. Why do you think it existed in the

ports industry when in most other industries workers are employed on a regular weekly basis?

2. Consider a group of n individuals each of whose production activities must be co-ordinated with the other $n - 1$ individuals. How many contracts are required under market co-ordination in which each individual contracts with every other individual? How many contracts are required if there is co-ordination via a central co-ordinator?
3. Set out as fully as possible the probable advantages and disadvantages of the producers' co-operative as compared to the conventional firm in the cases of:
 - (a) a group of six potters producing handmade pottery
 - (b) a group of two hundred workers producing motorcycles
 - (c) a group of four thousand workers producing a range of electrical and non-electrical components for motor cars.

C. Critique of the classical theory of the firm

If we compare a factual description of a modern firm with the abstract description implied by the 'orthodox' view of the firm set out at the end of the previous section, we find several marked differences. There appear to be many features of real firms which simply do not have counterparts in the theory:

- (a) *Ownership*: a firm may be owned by a single individual or by a small group, with each owner liable for the debts of the firm to the complete extent of his wealth. Alternatively, the firm may be owned by any number from a few to several thousands of people, with liability limited to the value of their ownership shares, exchangeable on a stock market. Part or all of the shares may be held by other firms, or financial organizations such as pension funds and insurance companies.
- (b) *Control*: where the firm is owned by one individual or a small group, it is likely that overall control will be exercised by someone with a significant ownership share. Where the ownership is dispersed over many individuals, overall control is exercised by a 'board of directors', acting in principle as fiduciary representatives of the owners, and comprising employees of the firm (senior executives) and 'outside directors'. Where the firm is partly owned by another firm or financial organization, some members of this group will often represent it on the board. If ownership is total, then control will usually be exercised by having the senior executives directly responsible to executives of the owning firm.
- (c) *Organization*: a hierarchical structure will exist between the people who directly carry out the basic activities of production, selling output and buying inputs, on the one hand, and the people exercising overall control, on the other. This is intended to fulfil a number of functions: to translate broad policy objectives formulated by controllers into specific plans; to co-ordinate the separate activities at lower levels and ensure consistency of plans;

to monitor performance, transmit information on this up to controllers and implement incentive systems; and to provide information with which overall policy objectives can be formulated. The larger the scale and greater the diversity of the basic production and selling activities of the organization, the more extended and complex this hierarchical structure will be.

(d) *Information*: the operations of the firm will generate information (reports from salesmen on demand conditions, performance of production processes, etc.) and also activities will be undertaken to acquire it (market research, technological research and development). This information must be transmitted to the points in the firm at which it is required for decision-making. The information will rarely be complete, so that decisions will generally be taken under varying degrees of uncertainty.

(e) *Conflict*: objectives, plans and decisions will generally be formulated or taken by more than one individual. Conflicts may arise between these individuals, for one or both of two reasons: because of lack of objective information, beliefs may differ about possible outcomes of decisions and the relative likelihoods of these; or preference orderings of the individuals over the outcomes of the decision may differ. The latter source of conflict is in turn due to the fact that a given outcome of a decision may benefit different decision-takers in different ways, so that conflict would be avoided only if they subordinated their own self-interest to some common objective, possibly that of the firm's shareholders. Moreover, although the direct participants in decision-taking are more often than not the executives of a firm, there will usually be other groups who can influence certain decisions by their behaviour. For example, workers can refuse to work if decisions about wages, hours and conditions of work do or do not take a certain form; shareholders can sell their shares if profits are low, and so on. The conflicts which exist among such groups will be reflected in decision-taking.

The description of the firm implicit in the profit maximization theory seems to include little of this. Nothing is said about control or organization structures and a very restricted view is taken of the nature of ownership. It seems to be assumed that whatever these may be, the firm will act in the best interests of its owners, i.e. the recipients of the 'residual income', and that there is no organizational problem in translating this objective into decisions. Moreover, no conflict is seen to exist: all decisions are taken in a way consistent with the objective of the firm. In all its decisions, including even inter-temporal ones, the firm has complete information, and there is no uncertainty. Thus, the theory clearly does not seem to take into account many features of reality.

However, any theory must abstract from or ignore *some* aspects of reality. The classical theory is designed to provide a foundation for models of the firm which allow us to derive relationships between output supply and input demand decisions, on the one hand, and changes in parameters such as output and input prices, taxes, and technological coefficients on the other. This can be done most tractably by taking a fairly abstract 'black box' model of the firm, and the real test of the value of this abstraction is the evaluation of how well it predicts the decisions firms take. There is no body of well-founded empirical evidence to support the view that, *in respect of the class of decisions with which it is concerned*, the classical theory gives false or misleading predictions. On the contrary the classical theory has proved its usefulness in many applications.

The more important and effective criticism is that the classical theory cannot help us analyse issues that arise out of the nature of the firm itself and which are of interest and importance quite independently of whether they affect qualitatively the forms of the firm's output supply and input demand relationships. The most important of these issues are

- (a) the consequences of the separation of ownership from control;
- (b) the 'boundaries' of the firm and the nature of the firm itself – if transactions costs are saved by integrating activities within a firm, why is the entire economy not one large firm?
- (c) the internal structure and organization of the firm – why hierarchy, and what is the most efficient 'architecture' for the structure of decision-taking within the firm?
- (d) the firm's capital structure, i.e. the appropriate mix of debt and equity;
- (e) the firm's 'internal labour market'.

The classical theory of the firm says nothing about these issues because it takes for granted the answers to the questions they pose. It would be wrong, however, to suppose that microeconomic theory has nothing to say on them. There has been a great deal of valuable work on these issues by economists working within the 'neo-classical' tradition, and they are still at the forefront of current research. In the remainder of this chapter, we shall consider briefly some approaches to the first two of these issues. More formal models are presented in Chapter 13.

Exercise 6C

1. Is the observation that, after a takeover, profits of the firm taken over are increased while the labour force and managerial staff are reduced consistent with the proposition that the objective of firms is to maximize profit?
2. Describe the ways in which (a) shareholders and (b) other firms acquire information about the profit-making possibilities of a particular firm.

D. Issues in the theory of the firm

The separation of ownership from control

As we saw in section B, the classical theory of the firm is based upon the idea of a central individual who: owns the firm's assets, which he finances by saving and borrowing; receives as his income the profits of the business; employs the inputs; bears the risks; and controls the firm. A different view of the nature of capitalism has become general, receiving impetus from the ever-increasing size of firms and concentration of economic activity. Though owner-controlled firms may still form the majority of business units, the bulk of economic

activity is controlled by large corporations, the distinguishing characteristic of which is a divorce of ownership from control. The owners of the firm are the shareholders, who may number thousands, and each of them has a very small proportion of the total equity of the company. They bear the risks, in the sense that fluctuations in the profits of the company imply fluctuations in their income – dividends – from it. They also supply the risk capital for new investment, either by buying new issues of shares, or, more usually, by 'agreeing' to forego profits which are then ploughed back into the business. On the other hand, by diversifying their shareholdings across a number of companies (or having experts do this for them, in unit trusts or investment trusts), they reduce the riskiness of the overall portfolio, until the main element of this may simply be the variability in general business conditions. A consequence of this is that shareholders take only a very general interest in the running of any one company. In exceptional circumstances, poor performance by a company could lead to 'stormy shareholders' meetings', and 'acrimonious debate', but the mechanisms of direct shareholder control are rarely effective. Rather, dissatisfied shareholders vote with their feet: they sell their shares, thus reducing the company's share price. This may represent an effective control on management, first because managers may like a high stock-market valuation (share price \times number of shares issued) for its own sake, but probably more importantly because a valuation which is low relative to the earning power of the assets of the company raises the threat of takeover by another company after which the present senior managers, at least, could lose their jobs. Hence there is some constraint on managers to take account of shareholders' interests.

In principle, the directors of a firm are meant to be the 'stewards' for the shareholders, to oversee the operations of the company on the shareholders' behalf. However, boards of directors are usually dominated by the senior executives of the company. Hence, if there were to be a conflict of interest between management and shareholders, it is far more likely that the board would espouse the interests of the former. Thus, the senior executives of the company rather than shareholders effectively control the decision-taking activities of the firm (subject to the organizational problems which they themselves may have in ensuring that their decisions are actually implemented).

The first thing deduced from this picture of *managerial capitalism* is that if managers' and shareholders' interests were in conflict, the divorce of ownership from control *permits* managers to pursue their own interests rather than those of shareholders, to an extent determined by the sanctions possessed by shareholders. The second step is to argue that the interests of managers and shareholders *do* conflict. It is suggested that managers derive their satisfactions from: their salaries; amounts of additional perquisites such as expense accounts, company cars, subsidized food and drink (which also have tax advantages); status, prestige, power and security. Although these things may depend on a profit performance which keeps shareholders happy, they do not necessarily vary directly with profit, but rather may vary with other dimensions of the firm's performance. To complete the argument, then, the third step is the proposition that, as we usually assume in economics, people take decisions in their own self-interest. This means that managers will choose output and input levels, and investment plans, in the light of their effects on the determinants of their own satisfactions, taking account of shareholders' interests only insofar as they represent an externally imposed constraint.

The initial attempts to construct models of managerial capitalism (by W. J. Baumol, R. Marris and O. E. Williamson) took the form of simple constrained maximization

problems. The managers' objectives are characterized by a single maximand such as sales revenue (Baumol) or growth (Marris), or by a utility function defined on variables which more directly reflect managerial preferences, such as expenditures on staff, salaries and 'perks' (Williamson). The interests and influence of shareholders are expressed by a minimum profit constraint, rationalized as the amount required to avoid takeover, and allowing the firm to make less than maximum profit. The models then provide a range of predictions about equilibrium choices and comparative statics responses which differ from those of the model of profit maximization, and also, in the case of Williamson's model, give interesting explanations of the existence of 'managerial slack' or 'X-inefficiency' (H. Liebenstein's term).

These models themselves beg a number of questions, however. The central weakness is the absence of any explicit analysis of the role of information and of the behaviour of the owner(s) of the firm – the simple assumption of a given profit constraint sweeps a lot of important issues under the carpet. A central implication of the separation of ownership from control is the *asymmetry of information* it creates between manager and owner. If an owner knew as much as the manager about the profit-making possibilities of the firm, the owner could threaten to punish observed deviations from profit maximizing behaviour and so ensure that they did not take place. On the other hand, if an owner perceives that the manager's information is inevitably going to be superior to his own, then he can try to devise a contract which takes account of this asymmetry of information and provides the manager with an incentive to take at least some account of the owner's objectives. For example, senior managers of large corporations are often given options to buy stock which encourage them to increase the value of the firm's shares. In addition, a large proportion of a senior manager's pay often consists of a profit-related bonus payment. We require an explicit analysis of the situation of asymmetric information to determine whether we can expect such incentive schemes to eliminate entirely the effects of the separation of ownership from control.

Relevant models are provided by *principal-agent theory*, which provides a general analysis of the following situation. A principal, P , employs an agent, A , to carry out some activity on her behalf. A must choose some decision variable, e , which determines an outcome $x = x(e, \theta)$. In this function θ is a random variable with known distribution. In one type of model, that of *moral hazard*, A must choose e before θ is known. P observes the outcome x , but cannot observe either e or θ and so is unable to ensure that A in fact chooses the value of e P would prefer. Her problem is then to design a contract that rewards A according to the outcome value x , taking into account any tendency A might have to choose a value of e which is non-optimal for P . In a second type of model, that of *adverse selection*, A knows θ before e is chosen, while again P cannot observe e or θ . But P knows that A knows θ , and so faces the problem of designing a contract which induces A to reveal the true value of θ . Clearly, if we identify P as the owner of the firm and A as the manager, principal-agent theory provides an important class of models with which to analyse explicitly the consequences of the separation of ownership from control. We set out some general principal-agent models in Chapter 22, and in Chapter 13 we apply them to the question of managerial incentives and the extent to which they can solve the problem of control.

Though principal-agent theory makes a significant contribution to our understanding of the consequences of the separation of ownership from control, it involves an important over-simplification. The principal is regarded as a single individual, whereas in reality there

may be a large number of owners of the firm. Under certain conditions this can be shown not to matter – shareholders would effectively be unanimous in their ranking of the decision alternatives open to the firm. However, there are realistic cases in which such unanimity could not be expected, and it would not be possible for a manager to act in the best interests of all shareholders even if he wanted to. This arises in particular when capital markets are *incomplete* – loosely, when the number of assets on the capital market, from which stockholders can construct asset portfolios, is less than the number of possible *states of the world*.

For example, suppose that next year the economy could either boom (state 1) or bust (state 2), and there is only one firm in the economy, with its shares as the only asset. The firm is owned by two individuals, Mr A and Ms B. It can carry out an investment which costs £100 and yields £110 if the economy booms and £80 if the economy busts. Mr A believes there is a probability of 0.8 of a boom, 0.2 of a bust. Ms B believes the probability of boom is the same as that of a bust, 0.5. For Mr A, the *expected value* of profit from the investment is $0.8(110) + 0.2(80) - 100 = £4$. For Ms B, the expected value of profit is $0.5(110) + 0.5(80) - 100 = -£5$. Should the manager of the firm undertake the investment? Clearly, no decision is in the best interests of both shareholders, as they would perceive them. Such conflicts can arise among shareholders not only because of differences in probability beliefs, but also because of different attitudes to risk. For example, suppose Ms B's probability beliefs are the same as Mr A's, but whereas he is prepared to accept an investment with positive expected value, she would regard the admittedly high chance of a profit of £10 not worth the risk of a loss of £20.

Now when capital markets are complete, this problem can be shown to disappear (see Chapter 21 for a full discussion of this). In effect, the capital market will provide a value for £1 in the state of the world 'boom', say r_1 , and a value for £1 in the state of the world 'bust', r_2 and both shareholders are better off if and only if $110r_1 + 80r_2 - 100 > 0$. Thus the manager has an objective criterion with which to act in his shareholders's best interests. However, if capital markets are incomplete these values do not exist, and we have the problem of how to formulate a decision criterion for the manager. We examine this problem in some depth in section 22 – below: at this point we simply note its relevance for the discussion of the implications of the separation of ownership from control. Managers may not act in the best interests of shareholders *not* because of a failure of control, but because those best interests are not well defined.

In a world of complete certainty and with perfect capital markets, the separation of ownership from control would have no consequences for profit maximizing behaviour (see Chapter 15 where this result is established). Since this is the kind of world for which the classical theory of the firm was developed, the theory is at least internally consistent. The development of the economics of uncertainty and information that has taken place over the last few decades now allows us to analyse cases in which the separation of ownership from control may have important consequences, and useful insights have been gained. There remain, however, unsolved problems, and the subject is firmly on the research agenda.

The nature and limits of the firm

As we noted earlier, in a seminal paper R. Coase argued that the firm's existence can be explained by the principle of economizing on transactions costs. In principle, production activities could be carried out by a system of contracts between owners of factors of

production, with these contracts being repeatedly re-negotiated in the light of prevailing market conditions. Although a firm can also be regarded as a complex of contracts – of employment, output supply, debt finance, and so on – activities are organized through a system of authority and control rather than through market transactions. We cannot claim to have understood the nature of the firm until we have understood why this is so. In addition, a satisfactory explanation of the limits of the firm must give reasons for the inclusion of some activities in, and the exclusion of others from, the set of activities made subject to direct organization within the firm.

The more recent work that has stemmed from Coase's insight, in particular that of O. E. Williamson, and S. Grossman and O. D. Hart, has concentrated on clarifying the nature of the 'transactions costs' which arise from using the market, and on applying the concepts thus developed to explain why some activities are more profitably integrated within the firm while others are left as market transactions. We shall proceed by first explaining these key concepts, and then showing how they fit together to give a theory of the nature and limits of the firm.

The key concepts are:

- (a) incompleteness of contracts
- (b) asset specificity
- (c) opportunistic behaviour
- (d) residual rights of decision and control.

We consider these in turn.

(a) *Incomplete contracts.* At the time a contract is concluded there is generally uncertainty about the future state of the world in which the provisions of the contract will actually be carried out. For example, a contract may specify delivery of a certain quantity of input at a specified price on a future date, and neither party will know for sure the market prices for that input and for the output it is to be used to produce, on that future date. As you will see in Chapter 19, one possible response would be to list every possible state of the world which might prevail on the future date – for example, every possible pair of values of the market prices of the output and input – and then make the terms of the contract contingent on these states of the world. Such a contract would be called *comprehensive* or *complete* (the sense of the word 'complete' here is related but not identical to that in which it was used in the above discussion of capital markets and the separation of ownership from control). However, contracts are rarely, if ever complete in this sense. One reason for this is that it would be too costly and time-consuming to draw up a list of all possible contingencies and negotiate terms for each of them. Another may be that whether or not some particular contingent event has occurred could not be verified by a third party, e.g. a court of law, and so performance of the contract could not be enforced. Therefore, *ex ante*, such performance could not be contracted for. Thus, for at least a subset of states of the world the terms of the contract could not be made state contingent and so the contract will be incomplete.

(b) *Asset specificity.* One or both parties may commit resources in the course of a trading relationship, and if the relationship should end and the resources are switched to another use, there would be significant loss of productive value or cost incurred.

Examples of such *relationship specific investments* are:

- (i) an electricity producer locates a power plant near a coal mine to minimize transport costs in using its output;
- (ii) a clothing manufacturer invests in special equipment required to produce the styles and quality of clothing specified by a large retailing chain;
- (iii) an individual invests time and money in acquiring particular skills which are of value only in the firm with which she contracts;
- (iv) an airport authority builds an air terminal specifically for the use of a single large airline.

In each case the assets created (in example (iii), *human capital*) have a value which is specific to a particular transaction.

(c) *Opportunistic behaviour.* The self-interest of economic agents will lead them to take whatever advantage they can of a given trading situation, even if it means suppressing or distorting information, behaving ruthlessly or unfairly, and so on. Business is business.

(d) *Residual rights of decision and control.* If a contract is incomplete, then it will not specify the actions of the parties for some states of the world. Ownership of the productive assets is then often associated with the power to decide what will be done in these circumstances, that is, how the assets are to be used. Ownership confers the *right to decide* in those situations not covered by a contract.

The deeper analysis of the 'transactions cost minimization' rationale for the firm developed primarily by O. E. Williamson, and further extended by S. Grossman and O. D. Hart, then proceeds as follows. The incompleteness of contracts *ex ante* creates scope for opportunistic behaviour *ex post*. Following the realization of some state of the world the parties may have to bargain and negotiate over the division of the returns from the activity, and each will exploit whatever advantages they may possess. This is particularly important when there is asset specificity; the owner of the relationship specific assets is at risk of being exploited, because of the costs of switching to alternatives. The awareness of this *ex ante* may cause underinvestment in such relationship specific assets. That is, an important form of the costs associated with market transactions where the parties are not part of the same firm, arises out of the distortions to investment in relationship specific assets caused by the awareness that contracts are incomplete and *ex post* opportunistic behaviour may reduce the returns to this investment. This cost may be avoided, if the transaction is organized within the firm. The owner of the assets will have the residual rights of control and this may change the *ex post* distribution of returns in such a way that the distortion to *ex ante* investment is corrected.

In exercise 6D we explore this rationale for the integration of activities within a firm in an example, which will bring out more clearly the roles played by the concepts just set out. Moreover, it will also bring out the fact that whether or not integration of an activity within a firm yields higher returns than leaving it to a market transaction carried out by means of an (incomplete) contract, depends on the values of certain key parameters which give formal expression to these concepts.

According to the 'transaction cost' theory the limits of the firm are determined by the set of activities which it is profitable to integrate within a structure in which the owner(s) of the assets (possibly delegating authority to managers) have the residual rights of control. This contrasts with, but does not exclude, the more traditional type of explanation, which sees the limits on size and scope of the firm as being determined by rising costs as the firm expands. Under this explanation, expansion of the firm's scale increases the size of the management hierarchy, causing increased bureaucracy and control loss. Moreover, we can view the internal organization of the firm as a complex system of principal-agent relationships, and the 'agency costs' (see section 13[C]) may increase as the hierarchy becomes larger. Interesting possible evidence on the existence of these kinds of inefficiencies is provided by the ability of 'corporate raiders' to buy up large companies, break them up and sell the constituent businesses separately, for a total sum greater than they paid for the conglomerate business as a whole.

Nevertheless, this traditional explanation of the limits to firm size, however plausible, suffers from two drawbacks. It does not really explain *which* activities will be integrated within the firm and which not. And it cannot answer Williamson's question concerning 'selective intervention'. It is always open to the owner of a conglomerate to structure it as a set of separate businesses (corporate divisions) and only intervene in their activities when it is clearly profitable to do so, so why then should the organizational diseconomies of scale arise? The evidence on takeovers may simply reflect the consequences of the separation of ownership from control – managers have not chosen the most profitable organization for the corporation. The transactions cost approach is able to characterize which activities will and which will not be integrated into the firm, in a way which does not fall foul of the selective intervention question.

Exercise 6D

- Supplier A can undertake an investment of 100 now. This investment will generate a net surplus of 200 if used in state of the world 1, and 120 if used in state of the world 2. The probabilities of these states are p and $1 - p$ respectively. The surplus is to be shared with the Distributor B . A contract can be concluded which will specify their shares, S_1 and $1 - S_1$, of the surplus in state 1, but no prior agreement in state 2, S_2 and $1 - S_2$, can be enforced. B must receive at least 20 in state 1. If the investment is not used in this activity it will be worthless. A will undertake the investment only if the expected value of her return, $pS_1 200 + (1 - p)S_2 120 > 100$.

- If B behaves opportunistically and A is rational, what should she expect to receive in state 2?
- Use this to define the set of values of p and S_1 such that A will *not* invest. Show that there is a largest probability $\hat{p} > 0$ such that A will not invest even if $S_1 = 1$.

- Explain why everyone is better off if A incorporates B into her firm when $0 < p \leq \hat{p}$.

E Conclusions

The purpose of this chapter was to give a general picture of the concept of 'the firm' in economic theory. Seven of the next eight chapters go on to develop the theory of the firm at some length. At many points, the analysis becomes detailed and technical, but, important as it is that the reader should master these details, the overall nature and purpose of the theory should not be lost sight of, and this chapter is designed to help in this. Six of the next seven chapters are concerned with developing the theory of the profit-maximizing firm: the optimization problem with which the firm is assumed to be confronted is examined in detail, under a range of assumptions about market conditions and technology, and a wide set of hypotheses about firm behaviour is derived. In Chapter 13 we consider some models of the firm which seek to explore more rigorously some of the newer ideas considered in this chapter.

References and further reading

Two fundamental papers on the question of why firms exist are

- R. Coase. 'The nature of the firm', *Economica*, 4, 1937, 386–405.
 A. Alchian and H. Demsetz. 'Production, information costs, and economic organization', *American Economic Review*, 62, 1972, 777–95.

There are good accounts of the more recent approaches to the theory of the firm based on incomplete contracts, relationship specific investments and principal-agent theory in

- B. R. Holmstrom and J. Tirole. 'The theory of the firm'.
 O. E. Williamson. 'Transactions costs economics',

both in

- R. Schmalensee and R. D. Willig (eds). *The Handbook of Industrial Organisation*, vol. 1, North Holland, Amsterdam, 1989

and

- S. Grossman and O. Hart. 'The costs and benefits of ownership: a theory of vertical and lateral integration', *Journal of Political Economy*, 94, 1986, 691–719.

CHAPTER 7

Production

The starting point for an analysis of the firm's production decision is the problem of minimizing the cost of producing a given level of output subject to technological constraints. This problem is an incomplete model of the firm because the level of output is taken as given, but it is still important for two reasons. First, minimization of production cost is a necessary condition for the maximization of the objective functions of several important models of the firm. Second, as we shall see in Chapter 17, least-cost production is a necessary condition for the efficient allocation of resources, and hence our results provide criteria for making judgements about the efficiency of resource allocations. In this chapter we examine in some detail the technological constraints in the firm's cost minimization problem, and leave the analysis of the problem itself to the next chapter.

A. The production function

The firm transforms a large number of different types of inputs into a number of outputs, but to simplify the analysis we initially consider the case of a firm using two inputs (z_1, z_2) to produce a single output y . In most of this chapter we use a *production function* to summarize the technical constraints on the firm's production decisions but in the final section we briefly introduce the *production possibility set* as an alternative description of the feasible output and input combinations.

We initially restrict y and the input vector $z = (z_1, z_2)$ to be non-negative. (In section D we show how suitable redefinition of variables makes this restriction unnecessary.) The firm's *production function* $f(z_1, z_2)$ shows the maximum output which can be produced from the input combination (z_1, z_2) so that the technological constraint on the firm's behaviour is

$$0 \leq y \leq f(z_1, z_2) = f(z) \quad [\text{A.1}]$$

If the firm's actual output from z is equal to the maximum feasible output it is *output efficient*. The possibility that the firm is output inefficient with $y < f(z)$ is allowed for primarily to investigate the circumstances in which the firm will *choose* to be output

efficient. In what follows we will often make the implicit assumption that the firm is output efficient and write the technological constraint on the firm as $y = f(z)$.

Different assumptions about technology imply different production functions with different properties. Since the appropriate assumptions may vary with the circumstances being modelled we do not attempt to lay down a set of axioms about technology to be satisfied in all cases. Rather we introduce a convenient set of concepts for describing the consequences of alternative assumptions about technology.

One obvious restriction that is imposed on the technology is that it is impossible to produce output without using any inputs:

$$f(0, 0) = 0 \quad [\text{A.2}]$$

There is *essentiality* if the production function satisfies [A.2]. More interestingly, if it is impossible to produce output without using a particular input, no matter how much of other inputs are used there is *strict essentiality*. For example, if input 1 is labour and labour is essential then

$$f(0, z_2) = 0 \quad [\text{A.3}]$$

For much of what follows we assume that $f(z)$ is twice continuously differentiable. Unlike [A.2] or [A.3] this is a strong assumption which may not be satisfied in many interesting cases. We make the assumption because it simplifies many definitions and derivations of results. The *marginal product* MP_i of input i in the production of y is the rate at which the maximum feasible output of y changes in response to an increase in z_i with the other input held constant. It is therefore the partial derivative of $f(z)$ with respect to z_i :

$$MP_i = \partial f_i(z_1, z_2) / \partial z_i = f_i(z)$$

We do not restrict the marginal product to be positive. For example, more fertilizer applied to a given amount of land will eventually reduce the crop. We do, however, make the plausible *productivity assumption* that there is always one input with a positive marginal product. Thus in the previous example if the marginal product of fertilizer is negative output can be increased by using the same amount of fertilizer on a bigger area of land.

The *input requirement set* $Z(y^0)$ for the output level y^0 is the set of input combinations which can produce at least y^0 :

$$Z(y^0) = \{z | f(z) \geq y^0\} \quad [\text{A.4}]$$

$Z(y^0)$ is the feasible set for the firm facing the problem of choosing z to minimize the cost of producing y^0 . It is clearly closed because of the weak inequality in [A.4]. If $Z(y^0)$ is convex then the firm's production function is quasi-concave. (Recall our discussion of utility functions and preferences in section 3A). In Fig. 7.1 the input requirement set for output level y^0 is the shaded area.

The *isoquant* $I(y^0)$ for the output level y^0 is the set of input combinations which can produce y^0 when used output efficiently:

$$I(y^0) = \{z | f(z) = y^0\} \quad [\text{A.5}]$$

The assumption that at least one marginal product is positive ensures that isoquants, like the indifference sets of section 3A, must be curves rather than areas. In Fig. 7.1 the isoquant

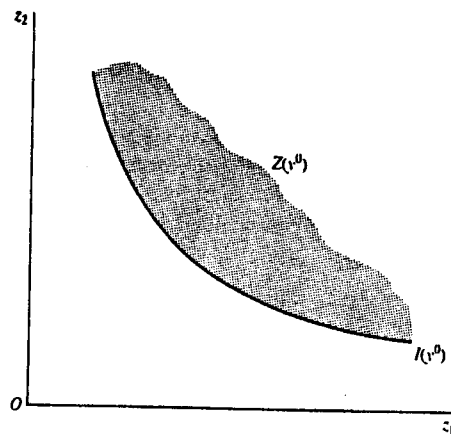


Fig. 7.1

for y^0 is the curve $I(y^0)$ which is the *boundary* of $Z(y^0)$. Those input combinations in the *interior* of $Z(y^0)$ can produce a larger output than y^0 and are therefore output inefficient. The combinations on the boundary of the input requirement set are output efficient for y^0 .

An isoquant is a *contour* of the production function since it satisfies the relation:

$$f(z) = y^0 \quad [\text{A.6}]$$

for some given y^0 . Differentiating this totally gives

$$f_1 dz_1 + f_2 dz_2 = dy^0 = 0$$

which can be rearranged to yield

$$-\frac{dz_2}{dz_1} \Big|_{dy=0} = \frac{f_1(z)}{f_2(z)} = \frac{MP_1}{MP_2} \quad [\text{A.7}]$$

The left-hand side of [A.7] is the negative of the slope of the isoquant and is the rate at which z_2 must be substituted for z_1 so as to keep output constant. It is the *marginal rate of technical substitution* of input 2 for input 1 and is denoted $MRTS_{21}$. It is directly analogous to the MRS_{21} of consumer theory. The utility function of consumer theory is an *ordinal* function whereas the production function involves a measure of output which is *cardinal* – the only degree of freedom in representing technology by a production function is in the choice of units of measurement of inputs or outputs. This gives the magnitude of marginal products f_i and their rates of change $f_{ij} = \partial f_i(z)/\partial z_j$, a significance which the magnitude of marginal utilities did not possess. Note that, just as the MRS_{21} was independent of the particular utility function chosen to represent preferences, the $MRTS_{21}$ is independent of the units in which output is measured.

In Fig. 7.2 I_0 , I_1 and I_2 are isoquants for successively greater output levels y^0 , y^1 and y^2 . (Note that a point on I_1 is in the interior of the input requirement set for y^0 and is therefore output inefficient for output y^0 and output efficient for output y^1 .) The negatively sloped segments of the isoquants arise when both inputs have positive marginal products (see [A.7]).

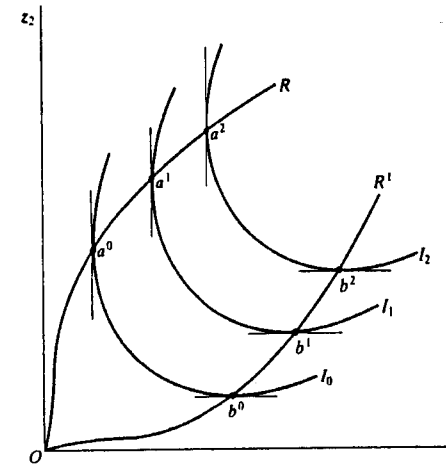


Fig. 7.2

Positively sloped portions of the isoquants occur when one of the inputs has a negative marginal product and the other a positive marginal product. For example, above a^0 on I_0 the marginal product of z_2 (say fertilizer) is negative and the reduction in output caused by further increases in z_2 must be offset by increases in the z_1 (say land) which has a positive marginal product.

At points like a^0 , a^1 and a^2 the marginal product of z_2 is zero and at points like b^0 , b^1 and b^2 the marginal product of z_1 is zero. The lines OR and OR' , connecting the points at which MP_2 and MP_1 respectively are zero, are *ridge lines*. The area inside the ridge lines is known as the *economic region* because a cost-minimizing firm would always choose a point within it. This can be seen easily in Fig. 7.2. For example, for every point on the isoquant I_2 (corresponding to output y^2) outside the economic region, there is a point on I_2 inside the region where less of both inputs is used to produce y^2 . Hence, as long as all inputs have non-negative prices and at least one has a positive price, a firm will incur a lower cost of producing y^2 inside the economic region than outside it.

The reason we bother to show the non-economic region outside the ridge lines is that it may be relevant in theories where the firm is not cost minimizing: it may, for example, have preferences which depend directly on the amount of an input used. Consideration of the non-economic region also leads to a distinction between output efficiency and *technical efficiency*. Production is technically inefficient if it is possible to produce a given output with less of at least one input and no more of another. Points on an isoquant for a given output level are output efficient but unless they are in the economic region they are not technically efficient.

Elasticity of substitution

As we will see in the next chapter the shape of the isoquants has important implications for the effect of a change in input prices on the input mix used to produce a given output.

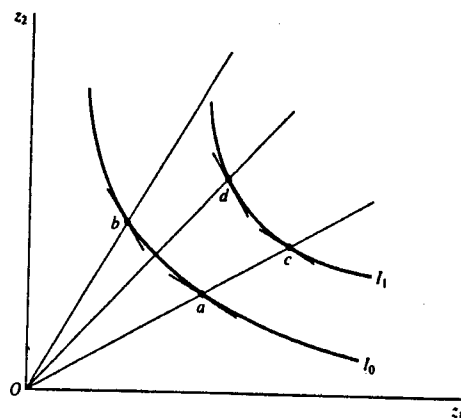


Fig. 7.3

In particular, we will be interested in the *elasticity of substitution*

$$\sigma = \frac{\% \text{ change in } z_2/z_1}{\% \text{ change in } \text{MRTS}_{21}} = \frac{d(z_2/z_1)}{d(f_1/f_2)} \cdot \frac{(f_1/f_2)}{(z_2/z_1)} \quad [\text{A.8}]$$

which captures the relationship between the input ratio and the curvature of the isoquants. Figure 7.3 illustrates. Consider the points a and b on the isoquant I_0 . The change in the output ratio between a and b is shown by the difference between the slopes of the rays Ob and Oa . The corresponding change in the MRTS_{21} is shown by the difference in the slopes of the lines tangent to I_0 at b and a . Now consider the isoquant I_1 and the points c and d . The slope of I_1 at c and d is equal to slope of I_0 at a and b respectively. The input mix is the same at c and a but the ratio z_2/z_1 is smaller at d than at b . Thus I_1 has a smaller elasticity of substitution than I_0 : a smaller proportionate change in the input mix is associated with the same proportionate change in the slope of the isoquant. Intuitively: the smaller is the elasticity of substitution the more 'bowed in' will be the isoquants and the smaller the proportionate change in the input mix associated with any given proportionate change in the slope of the isoquant.

Exercise 7A

1. Explain what is meant by (a) technically efficient input combination and (b) an output-efficient input combination and show that for an input combination to be technically efficient it is necessary but not sufficient that it be output-efficient.
2. Why can isoquants have positively sloped regions while indifference curves do not?
3. Why can we adopt an assumption of diminishing marginal products when we could not adopt an assumption of diminishing marginal utility?

4. Fixed Proportions Technology (Leontief)

- (a) Process 1 uses at least β_{11} units of z_1 and β_{12} units of z_2 to produce one unit of output. Draw the isoquant for $y = 1$, and distinguish between the technical and output efficient (z_1, z_2) points. Suppose that at least $y \cdot \beta_{11}$ units of z_1 and $y \cdot \beta_{12}$ units of z_2 are required to produce y units of output, so that the production function for process 1 is

$$y = \min \left(\frac{z_1}{\beta_{11}}, \frac{z_2}{\beta_{12}} \right)$$

Draw the isoquant map for the process. What does the economic region look like? (Note: $\min(\cdot, \cdot)$ is read: 'the smaller of' the terms in brackets.)

- (b) Suppose that y can also be produced from process 2, which requires at least $y \cdot \beta_{21}$ of z_1 and $y \cdot \beta_{22}$ of z_2 and that processes 1 and 2 are *additive* in that the output from one process is independent of the level at which the other process is used. Under what circumstances would it never be technically efficient to use process 2? A given level of y could be produced by different mixtures of the two processes using different total amounts of the inputs. Derive the isoquant for mixtures of the two processes (where a mixture uses $k\beta_{11} + (1-k)\beta_{21}$ of z_1 and $k\beta_{12} + (1-k)\beta_{22}$ of z_2 to produce 1 unit of output, with $0 \leq k \leq 1$). A mixture is a convex combination of processes. (Compare the analysis of the Lancaster consumption technology model in section 5B.)
- (c) Let there be three, four, ..., n processes satisfying the above assumptions. Investigate the circumstances in which particular processes are never used. Show that as the number n of technically efficient processes becomes large the isoquant tends to the smooth shape assumed in section A.

5. The Cobb-Douglas production function is

$$y = \alpha z_1^\alpha z_2^\beta \quad \alpha > 0 \quad \beta > 0$$

Show that $MP_1 = \alpha y/z_1$, $MP_2 = \beta y/z_2$

What is the MRTS_{21} ? How does it vary with: (a) y ; (b) z_2/z_1 ?

Draw the isoquant map.

6.* The CES production function is

$$y = A(\delta_1 z_1^\rho + \delta_2 z_2^\rho)^{1/\rho}, \quad \delta_1 + \delta_2 = 1, \quad A > 0$$

Show that

$$MP_1 = A^\rho \delta_1 \left[\frac{y}{z_1} \right]^{1-\rho}$$

What is the MRTS_{21} ? How does it vary with: (a) y (b) z_2/z_1 ?

7.* Elasticity of substitution.

- (a) Show that the elasticity of substitution can be written as

$$\sigma = \frac{f_1 f_2 (z_1 f_1 + z_2 f_2)}{z_1 z_2 [2f_{12} f_1 f_2 - f_{11} (f_2)^2 - f_{22} (f_1)^2]}$$

(Hint: define the input ratio as $r = z_2/z_1$, use the definition of the isoquant as $y^0 - f(z_1, rz_1) = 0$ to get z_1 as function of r : $z_1 = g(r)$ and write

$MRTS_{21} = f_1(g(r), rg(r))/f_2(g(r), rg(r))$. Differentiate $MRTS_{21}$ with respect to r (using the implicit function theorem to get dg/dr .)

- (b) Show that the elasticities of substitution of the Leontief, Cobb–Douglas and CES production functions are respectively zero, 1 and $1/(1 - \alpha)$. (Hint: in the latter two cases, rather than using the above expression for σ , use your earlier results concerning the MP_i and thus the $MRTS_{21}$ and remember the relationship between elasticities and logs.)

- (c) Explain why the elasticity of substitution of the linear production function

$$y = a_1 z_1 + a_2 z_2, \quad a_1 > 0, \quad a_2 > 0$$

is infinite. (Sketch the isoquants.)

8.* Show that the CES production function reduces to

- (a) the linear production function ($\alpha = 1$);
 (b) the Cobb–Douglas production function (as $\alpha \rightarrow 0$);
 (c) the Leontief production function (as $\alpha \rightarrow -\infty$).

B. Variations in scale

In this and the next section we examine further the responses of output to changes in inputs. The sections are a preparation for the investigation in Chapter 8 of the relationship between output and cost minimizing input choices. Since cost minimization implies technical efficiency we restrict attention in these sections to firms which are output efficient and operating in the economic region of the production function.

Changes in output can arise from

- (a) changes in the scale of production by varying all inputs in the same proportion; or
 (b) changes in relative input proportions.

The first corresponds to movements along a ray through the origin, such as OA or OB in Fig. 7.4, the second to a movement from one ray to another. For example, output can be increased by moving from z^0 on I_0 to the higher I_2 isoquant, either by doubling both inputs (moving to z^2), or varying the input proportion and moving to z^3 where z_2/z_1 has fallen. In this section we consider variations in scale and in the next an important case of variations in input proportions resulting from varying one input with the other held constant.

Starting from say z^0 on I_0 in Fig. 7.4 and multiplying each input by the scale parameter $s \geq 0$ is equivalent to a movement along the ray OA through z^0 . If $s < 1$ the scale of production is reduced and there is a movement toward the origin. Conversely if $s > 1$ the scale of production is increased and there is a movement away from the origin. For example, when $s = \frac{1}{2}$ the point z^1 is reached and when $s = 2$ the point z^2 is reached.

When we investigate the effects of scale variations from some initial input combination z we can write the production function as

$$y = f(sz) = y(s; z)$$

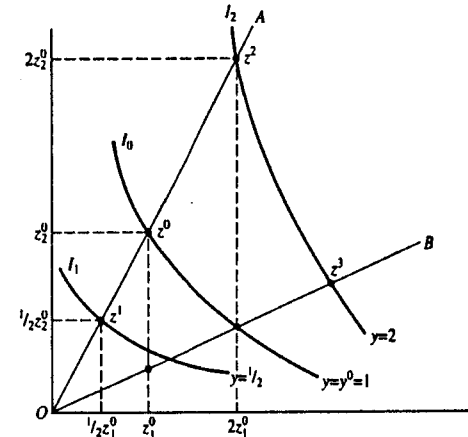


Fig. 7.4

and consider how y varies with the scale parameter s with input proportions held constant at the values implied by the initial z .

The elasticity of scale E is the proportionate change in output y divided by the proportionate change in the scale of production s :

$$E = \frac{dy}{y} \cdot \frac{s}{ds} = \frac{dy}{ds} \cdot \frac{s}{y} \quad [\text{B.1}]$$

It is a measure of the responsiveness of output to equal proportionate changes in all inputs. Output increases more or less proportionately with scale as E is greater or less than 1. There are said to be *increasing*, *constant* or *decreasing returns to scale* as $E > 1$, $E = 1$, or $E < 1$. Since dy/ds depends on the input mix as well as the scale parameter the returns to scale for a production function may depend on the input mix and the scale. Thus in Fig. 7.4 examination of the I_0 and I_2 isoquants shows that there are constant returns along OA and increasing returns along OB . In Fig. 7.5 output is plotted against the scale parameter (so that input proportions are held constant) and a number of possibilities are illustrated. There are increasing returns in part (a), constant returns in part (b), decreasing returns in part (c) and in part (d) there are initially increasing and then decreasing returns to scale.

Homogeneous and homothetic production functions

A production function is *homogeneous of degree t* if multiplying all inputs by the scale parameter s causes output to increase by the factor s^t . Formally if

$$f(sz) = s^t f(z) \quad [\text{B.2}]$$

then the production function $y = f(z)$ is homogeneous of degree t . When $t = 1$ the production function is *linear homogeneous*. Many models assume that $f(z)$ is linear

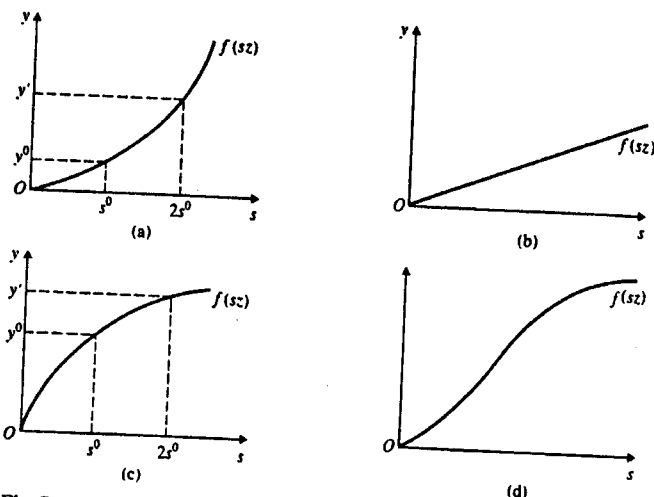


Fig. 7.5

homogeneous because such functions have a number of properties which greatly aid analysis.

(a) Using [B.2] the elasticity of scale of a homogeneous function of degree t is

$$E = \frac{dy}{ds} \frac{s}{y} = \frac{df(sz)}{ds} \cdot \frac{s}{f(sz)} = \frac{ds'f(z)}{ds} \cdot \frac{s}{s'f(z)} = ts^{t-1}f(z) \frac{s}{s'f(z)} = t$$

Since a linear homogeneous function has $t = 1$ we see that the linear homogeneous production function has constant returns to scale at all input combinations.

(b) [B.2] must hold for all z and so the partial derivatives with respect to z_i of both sides of [B.2] must be equal. Since the partial derivative of the left hand side is

$$\frac{\partial f(sz_1, sz_2)}{\partial z_i} = \frac{\partial f(sz_1, sz_2)}{\partial (sz_i)} \cdot \frac{d(sz_i)}{dz_i} = f_i(sz)s$$

and of the right-hand side is $s'f_i(z)$ we have $f_i(sz)s = s'f_i(z)$ or

$$f_i(sz) = s^{t-1}f_i(z)$$

[B.3]

Hence a function which is homogeneous of degree t has partial derivatives which are homogeneous of degree $t - 1$. Since $t = 1$ for a linear homogeneous function we have established that $f_i(sz) = f_i(z)$: linear homogeneous production functions have marginal products which are independent of scale. The marginal products will depend only on the input proportions and will be constant along rays from the origin.

(c) The slope of the isoquant at sz is $-f_1(sz)/f_2(sz)$ and at z is $-f_1(z)/f_2(z)$. Since [B.3] holds for all inputs i when the production function is homogeneous of degree t we see that the slopes of the isoquants of a homogeneous production function depend only on the input proportions and are independent of the scale of production. The slopes will be constant along rays from the origin and each isoquant is a radial expansion or contraction

every other isoquant. As we show in the next chapter this implies that input proportions for cost minimizing firms depend only on input prices and not on the level of output.

(d) Since [B.2] holds for all s if f is homogeneous the derivatives of both sides of [B.2] with respect to s must be equal. The derivative of the left-hand side of [B.2] is

$$\frac{df(sz_1, sz_2)}{ds} = \sum_i \frac{\partial f(sz_1, sz_2)}{\partial (sz_i)} \cdot \frac{d(sz_i)}{ds} = \sum_i f_i(sz)z_i = s^{t-1} \sum_i f_i(z)z_i$$

where the last step follows from [B.3]. The derivative of the right-hand side of [B.2] is $s^{t-1}f(z)$ and so when f is homogeneous of degree t

$$\sum_i f_i(z)z_i = tf(z) \quad [B.4]$$

This result is known as *Euler's Theorem*. When the production function is linear homogeneous [B.4] gives the *adding up property*: output is equal to the sum of the marginal products of the inputs times their level of use. Its significance is that if the price of each input is equal to the value of its marginal product (the price of output times the marginal product) then a profit maximizing firm will break even: its revenue will be equal to its costs.

A production function $g(z)$ is *homothetic* if it can be written as an increasing transformation of a linear homogeneous function of the inputs: $g(z) = F(f(z))$ where $f(z)$ is linear homogeneous, $F' > 0$ and $F(0) = 0$. Think of $f(z)$ as a 'composite' input and of $F(f)$ as a single input production function. An example of a homothetic production function is

$$y = \ln(z_1^\alpha z_2^{1-\alpha}) = \alpha \ln z_1 + (1-\alpha) \ln z_2, \quad 0 < \alpha < 1 \quad [B.5]$$

where $F = \ln f$ and f is just a constant returns Cobb-Douglas production function.

Homothetic production functions are important because, unlike homogeneous functions, they can have variable returns to scale but they also have the useful property (c) of homogeneous functions. As the reader should check, in the example [B.5] of a homothetic production function, increasing all inputs by the factor s will increase output by the factor $\ln s$.

The elasticity of scale of the homothetic production function is

$$E = \frac{dg(sz)}{ds} \cdot \frac{s}{g(sz)} = \frac{dF(f(sz))}{df} \cdot \frac{df(sz)}{ds} \cdot \frac{s}{F(f(sz))} = \frac{dF}{df} \cdot \frac{df}{ds} \cdot \frac{s}{f} \cdot \frac{f}{F} = \frac{dF}{df} \cdot \frac{f}{F}$$

(remember that from the definition of homotheticity f is linear homogeneous so that $df/ds \cdot s/f = 1$). Thus the scale elasticity of a homothetic production function $F(f(z))$ is not constrained by the requirement that f is linear homogeneous. The scale elasticity of [B.5] for example is $1/F$ which decreases with the scale of production.

The marginal product of input i if g is homothetic is

$$g_i(z) = F'(f(z))f_i(z)$$

and so the slope of the isoquant is

$$-g_1(z)/g_2(z) = -F'(f) f_1(z)/F'(f) f_2(z) = -f_1(z)/f_2(z)$$

Since f is linear homogeneous $-f_1/f_2$ depends only on the relative input proportions (property (c)) and so the slopes of the isoquants of a homothetic production function are independent of scale.

Exercise 7B

1. Do all homogeneous production functions of whatever degree have (a) marginal products and (b) marginal technical rates of substitution which are independent of the level of output?
2. What are the degrees of returns to scale for the (a) linear production function; (b) Leontief production function; (c) Cobb–Douglas production function; (d) CES production function?
3. Show that all homogeneous functions are also homothetic. Give an example (not [B.5]) of a production function which is homothetic but not homogeneous of any degree.
4. *Elasticity of substitution and linear homogeneity.* Show that if $f(z_1, z_2)$ is linear homogeneous the expression for the elasticity of substitution in Question 7, of exercise 7A simplifies to $\sigma = f_1 f_2 / y f_{12}$. (Hint: use the fact that [B.4] implies $f_{11} z_1 + f_{22} z_2 = 0$ to substitute for f_{11} and f_{22} .)

C. Variations in input proportions

Figure 7.6 illustrates the effects of changes in input proportions when one input (z_2 in this case) is held fixed and the other is free to vary. In part (a) the isoquant map is shown and z_2 is assumed fixed at z_2^0 . Variations in z_1 will lead to a movement along the line through z_2^0 parallel to the z_1 axis, and the output of y produced with $z_2 = z_2^0$ for different levels of z_1 can be read off from the isoquants. Part (b) plots the total curve $y = f(z_1, z_2^0)$ which results. If part (a) can be thought of as the contour map of the total product hill then part (b) shows a vertical slice through the hill at $z_2 = z_2^0$. Holding z_2 at different levels will give rise to different total product curves. Part (c) shows the average and marginal product of z_1 as a function of z_1 and is in turn derived from the total product curve of part (b).

The average product of z_1 , $AP_1(z_1, z_2^0)$ is total product divided by z_1 : y/z_1 . Consider in part (b) a ray from the origin to a point on the total product curve, for example the line OB . The slope of this line is the vertical distance BC divided by the horizontal distance OC . But $BC = y^0$ and $OC = z_1$ and hence: slope $OB = BC/OC = y^0/z_1 = AP_1(z_1, z_2^0)$. The AP_1 curve is, therefore, derived by plotting the slope of a ray from the origin to each point on the total product curve.

The marginal product curve MP_1 is derived by plotting the slope of the total product curve. Notice the relationship between the AP_1 and MP_1 , with the MP_1 cutting the AP_1 from above at the point z_1^* where AP_1 is at a maximum. It can be demonstrated that this relationship is no accident of draughtsmanship. The definition of the average product is:

$$AP_1 = \frac{y}{z_1} = \frac{f(z_1, z_2^0)}{z_1} \quad [\text{C.1}]$$

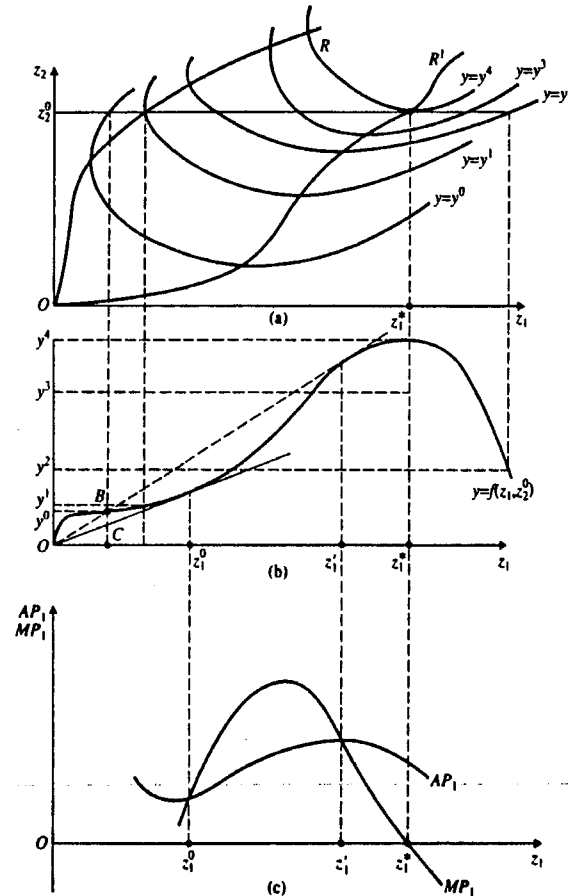


Fig. 7.6

Differentiating and setting equal to zero as a necessary condition for maximization yields

$$\frac{d}{dz_1} [AP_1] = \frac{1}{(z_1)^2} \left[\frac{\partial f}{\partial z_1} \cdot z_1 - f \right] = 0 \quad [\text{C.2}]$$

$$= \frac{1}{z_1} \left[f_1 - \frac{f}{z_1} \right] = \frac{1}{z_1} [MP_1 - AP_1] = 0 \quad [\text{C.3}]$$

Hence $MP_1 = AP_1$ is a necessary condition for AP_1 to be maximized.

Exercise 7C

1. What is the significance of the fact that in Fig. 7.6, the input level z_1^* is at the same time (a) a co-ordinate of a point on the ridge line, (b) the value of z_1 at which y is a maximum given $z_2 = z_2^0$, (c) the value of z_1 at which MP_1 is zero?

2. Explain why, in Fig. 7.6, AP_1 is at a minimum and $MP_1 = AP_1$ at z_1^0 .
3. Redraw Fig. 7.6 taking a fixed level of z_1 rather than z_2 .

D. The multi-product case*

In the previous sections we have written the firm's production function in the *explicit* form $y = f(z_1, z_2)$, or, allowing for output inefficiency, $y \leq f(z_1, z_2)$. When the firm produces more than one output it is often more convenient to write the production function in its *implicit* form. Corresponding to the two explicit cases above we could have

$$y - f(z_1, z_2) = g(z_1, z_2, y) = 0$$

or

$$y - f(z_1, z_2) = g(z_1, z_2, y) \leq 0$$

The implicit and explicit forms are equivalent ways of describing the technical constraints on production provided that we restrict attention to the economic region of the explicit production function. (When isoquants are positively sloped the implicit production function is not well defined because given y and say z_2 there are two values of z_1 which satisfy $y - f(z) = 0$.)

The marginal products of the inputs and the marginal rate of technical substitution between them are derived from the implicit form by the implicit function rule of differentiation. Applying the rule we have for example

$$\frac{dy}{dz_1} = \frac{-g_{z_1}}{g_y} = -\left(\frac{-f_1}{1}\right) = f_1 = MP_1$$

It is also convenient in many cases to adopt a slightly different notational convention. We have so far talked of y as an output and z_1 as an input and restricted both outputs and inputs to being non-negative. But what is an input for one firm may be an output for another, or a firm may change from producing a good to using it as an input, or it may use part of an output as an input (a power station uses electricity for lighting in producing electricity). To save relabelling the good when this happens it is easier to use the concept of the firm's *net output* of a good. (Note: This is quite different from the meaning of the term net output as it occurs, say, in national income accounting, namely as the difference between a firm's revenue and cost of bought-in inputs.) If the net output is positive the firm is producing the good, if it is negative the firm is 'consuming' it or using it as an input. The firm's net output of good i will be written as the variable y_i which is not constrained to be non-negative. If $y_i > 0$ good i is produced or supplied by the firm, if $y_i < 0$ good i is 'consumed' by the firm and if $y_i = 0$ the good is neither produced nor consumed. Using this labelling we can rewrite the implicit production constraint in the general case as

$$g(y_1, y_2, \dots, y_n) = g(y) \leq 0 \quad [\text{D.1}]$$

with $y = (y_1, y_2, \dots, y_n)$ now defined as the net output vector. The y_i are sometimes referred to as *netputs*.

When $g(y) < 0$ there is output inefficiency and when $g(y) = 0$ there is output efficiency. A technically infeasible net output or netput vector is indicated by $g(y) > 0$. Notice that an *increase* in y_i means that if i is an input ($y_i < 0$) the use of the input has been *reduced*: y_i is measured along the negative part of the relevant axis if it is an input. When $g(y) = 0$ it is technically infeasible to increase an output or reduce an input without reducing the level of some other net output, i.e. reducing some other output or increasing some other input. This implies that the partial derivatives g_i are always positive at $g(y) = 0$ to reflect the fact that *ceteris paribus* increases in y_i (reducing an input or increasing an input) are not technically feasible because they would lead to $g > 0$.

Again using the implicit function rule on $g(y) = 0$ and allowing only y_i and y_j to change, we have:

$$\frac{dy_i}{dy_j} = \frac{-g_j}{g_i} \quad i, j = 1, 2, \dots, n \quad [\text{D.2}]$$

and this can be given a number of interpretations depending on whether y_i and y_j are positive or negative:

(a) $y_i < 0$ and $y_j < 0$. Both goods are inputs so that dy_i/dy_j is the rate at which one input can be substituted for another when all other goods (inputs and outputs) are held constant. It is therefore the (negative of) the marginal rate of technical substitution, i.e. it is the slope of the isoquant, which in the multi-product case is the boundary of the set of y_i, y_j combinations which will just produce a given level of the firm's outputs with all other inputs held constant. For example, in the single output-two input case considered in previous sections we have (remembering that an *increase* in y_i means a *decrease* in z_i):

$$\frac{dy_2}{dy_1} = -\frac{g_1}{g_2} = \frac{-f_1}{f_2} = -\text{MRTS}_{21} \quad [\text{D.3}]$$

Figure 7.7(a) shows the isoquant for given levels of y_3, \dots, y_n for a particular production function which has the convexity and smoothness properties of the explicit function of previous sections. Again all points in the shaded area are technically possible ($g(y) \leq 0$) but only points on the boundary of it are output efficient ($g(y) = 0$).

(b) $y_i > 0, y_j < 0$. Good i is an output, j an input, so that dy_i/dy_j is the rate at which the output of i changes when input j is reduced with all other outputs and inputs held

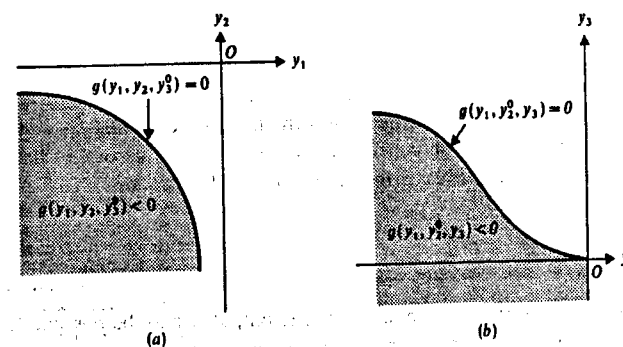


Fig. 7.7

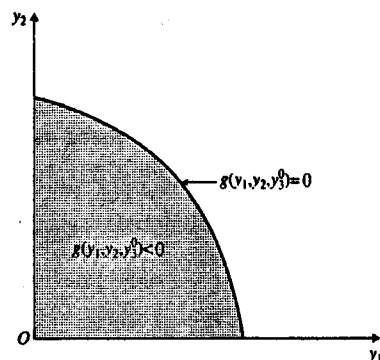


Fig. 7.8

constant. It is therefore the negative of the marginal product of input j in the production of output i . Using our single output, two input example and remembering that $y_i = -z_i$

$$\frac{dy_3}{dy_1} = \frac{-g_1}{g_3} = -f_1 = -MP_1 \quad [\text{D.4}]$$

Figure 7.7(b) shows the relationship between an input (good 1) and an output (good 3) and corresponds to the total product curve of Fig. 7.6(b). All points in the shaded area are technically possible but only points on the upper boundary (the total product curve) are output efficient.

(c) $y_i > 0, y_j > 0$. Both goods are outputs and so dy_i/dy_j is the rate at which the output of i varies as the output of j is increased when all inputs and all other outputs are held constant. This is the negative of the *marginal rate of transformation* of i into j or MRT_{ij} . In Fig. 7.8 both goods 1 and 2 are outputs, and the shaded area is the set of all technically possible combinations. The upper boundary of this shaded area is the set of output efficient points and is known as the *transformation curve*. Its slope is the marginal rate of transformation.

Increases in y_j require reductions in y_i . Different transformation curves are generated by fixing the other net outputs at different levels. A reduction in any other net output shifts the transformation curve out from the origin. In other words, decreases in other outputs or increases in inputs allow more of both good 1 and 2 to be produced. We have assumed that the technology allows substitutability of outputs so that the transformation curve is negatively sloped. If the outputs must be produced in fixed proportions (as, for example, in some chemical processes) the transformation curves would be rectangular, indicating that an increase in the output of one good requires an increase in the level of inputs and cannot be made by reducing the output of the other good.

Joint products

In some cases where a firm produces more than one output it may be possible to relate the output of each product to a specific part of the bundle of inputs used by the firm, so

that the firm has a production function for each output. For example if y_1 and y_2 are the levels of the firm's outputs and z_i is the amount of input i used in production of good j the firm's production possibilities could be written explicitly as

$$\begin{aligned} y_1 &\leq f^1(z_1^1, z_2^1) \\ y_2 &\leq f^2(z_1^2, z_2^2) \end{aligned} \quad [\text{D.5}]$$

or implicitly as

$$\begin{aligned} y_1 - f^1(z_1^1, z_2^1) &\leq 0 \\ y_2 - f^2(z_1^2, z_2^2) &\leq 0 \end{aligned} \quad [\text{D.6}]$$

When it is possible to describe the technical constraints on the firm in this way the production function is *separable*. If the firm is producing several products and inputs *cannot* be assigned in this way the firm is said to be producing *joint products*. Notice that it is the way in which the inputs relate to outputs, *not* the number of products, which is the defining characteristic of joint production. When the production function is separable the firm could be regarded as the sum of several single-product plants, and if each of the constituent plants acts to minimize the cost of its own production, total costs are minimized. Production can be *decentralized* without increasing cost. When there is joint production decentralization (instructing each product division to minimize cost) will not lead to minimum total cost because of the interdependence between the costs of each product. This point will be elaborated in the exercises in Chapter 9.

The production possibility set*

An alternative and more general way of describing the technological constraints on the firm is by its *production set*, PS , which is the set of all possible input-output combinations. The PS is the set of all feasible net output bundles, or of all feasible *activities*. An *activity* of the firm is the firm's net output bundle: $y = (y_1, \dots, y_n)$. The production function $g(y) \leq 0$ and the PS are equivalent descriptions of the technological constraints in the sense that the statement that y^0 is in the PS is equivalent to the statement $g(y^0) \leq 0$. If the activity y^0 is not technically possible then it is not in the PS and $g(y^0) > 0$. In terms of the figures in this section, the shaded areas (including their boundaries) can be thought of as slices through the PS and all points in the shaded areas are in the PS . The upper boundary of the PS is the set of points with the property that it is not possible to increase the net output of any good without reducing the net output of some other good (i.e. reducing an output or increasing an input). This upper boundary is therefore output-efficient and satisfies the equation $g(y) = 0$.

References and further reading

Good introductions to production theory are

- R. G. Chambers. *Applied Production Analysis: A Dual Approach*, Cambridge University Press, Cambridge, 1988, chs 1, 6.
- C. E. Ferguson. *The Neo-Classical Theory of Production and Distribution*, Cambridge University Press, Cambridge, 1969, chs 1-6.

There is a rigorous account in

R. W. Shephard. *Theory of Cost and Production Functions*, Princeton University Press, Princeton, NJ, 1970, chs 1-3.

The classic reference on the linear programming approach to production is

R. Dorfman, P. A. Samuelson and R. Solow. *Linear Programming and Economic Analysis*, McGraw-Hill, London, 1958, chs 1, 6, 9, 10.

There is a clear treatment of homogeneous functions and the elasticity of substitution in

R. G. D. Allen. *Mathematical Analysis for Economists*, Macmillan, London, 1938, chs 12-14.

CHAPTER 8

Cost

A. Introduction

Time dimension of production

In the previous chapter we did not consider in any detail exactly what is meant by 'inputs' and 'outputs' and in particular we did not discuss the time dimension of the firm's production function, preferring instead to talk loosely of 'levels' of outputs and inputs, in order to concentrate on the technical relationships involved. Output, however, is a *flow* and so must always have a time dimension: it is meaningless to say that a firm produces so many tons of a particular good unless we also specify the period of time (hour, day, month or year) over which the output was produced. y therefore has the dimension of a rate of flow of units of the good *per unit of time* or *per period*.

Input levels must be similarly interpreted. This is straightforward with inputs such as raw materials which are transformed or consumed by the firm. z_i would then have the dimension of the flow of the quantity of raw material of type i per period. Durable assets, however, such as machines, are, as the term implies, not consumed by the firm. In these cases we can think of the asset itself as embodying a *stock of productive services* and z_i is the *flow of productive services* of the asset used per period of time. For example with a machine of type i , z_i would be machine hours (the number of hours the machine is used) per day. The *capacity* of an asset is the maximum possible flow of productive services which can be used per period. In the example above the capacity of the machine is 24 machine-hours per day (assuming no time has to be taken for cooling down, maintenance, etc.). As we will see in section C it is often necessary to distinguish carefully between capacity and actual usage.

In this chapter, an 'input' will always be measured as a rate of flow, either of some physical good (coal, crude oil, cotton) or of the *services* of some factor of production which is not itself used up in the production process (labour, machinery).

Long- and short-run decision-making

We concentrate on a two-input model and we assume that z_1 is a *variable input*: it can be varied at will by the firm. The firm can decide at the start of period 0, the 'present' time period, to use any level of z_1 in production in period 0 and can implement that decision. The other input z_2 takes time to vary: it takes one period to make available an increment of z_2 , for example the flow of services from a machine or type of skilled labour. A decision taken 'now' at the start of period 0 to increase the amount of z_2 by Δz_2 will result in that increment becoming available for use in producing y at the start of period 1. As far as production in period 0 is concerned z_2 is a *constrained input*. The amount of z_2 used in period 0 certainly cannot be increased beyond the amount available at the start of period 0. On the other hand, the firm may or may not be able to *reduce* the amount of z_2 it uses in period 0. If the input is divisible the firm will be able to use less than the maximum amount unless there is some contractual limitation. Since contracts usually stipulate the amount of an input which will be *paid for* rather than the amount which must be used, divisibility will usually imply the possibility of using an input below capacity. For example, a firm may hire labour on a monthly contract, and be unable to increase or reduce the number of workers to whom it must pay a guaranteed weekly wage within that period, but it may *if it chooses* use less than the maximum possible number of man-hours.

The distinction between fixed and variable inputs has a crucial implication for the firm's decision-making. The firm is always located in time at the start of period 0 and at that moment of time it must make two types of decision. First, given the desired output level for period 0, it must choose an actual level of z_1 for period 0, remembering that maximum z_2 is fixed in period 0. (When z_2 can be less than its maximum level the firm must also choose an *actual* level of z_2 to be used in period 0.) Second, given the planned or desired output level for period 1, it must *formulate a plan* specifying desired levels of z_1 and z_2 to be used in period 1. If the desired amount of z_2 in period 1 in the plan differs from the level of z_2 held by the firm at the start of period 0 it must begin to organize the required change at the start of period 0, so that it is available at the start of period 1. Thus the choices *implemented* by the firm in period 0 are the input levels actually used in period 0, and the change in the fixed input available for the next period.

To predict how the firm's behaviour will vary in response to changes in the desired output levels in periods 0 and 1 or changes in the costs of inputs, we must construct a model of the two kinds of decision taken by the firm at the start of period 0. In section B we consider the problem of finding desired levels of z_1 and z_2 to minimize the cost of producing the planned period 1 output. Both inputs are variable in this problem since the firm will be able to bring about any planned change in z_2 by the start of period 1. This is referred to as the *long-run* cost minimization problem. In section C we model the problem of setting z_1 with a fixed maximum z_2 , so as to minimize the cost of producing the required period 0 output. This is the *short-run* cost minimization problem.

Adjustment costs

We assumed above that it was impossible to increase z_2 within period 0 but that z_1 was freely variable. This distinction is a crude recognition of the fact that in general there are

differing *adjustment costs* for different types of inputs. Adjustment costs are those costs which arise solely from a change in the level of use of an input. For example, if a firm wishes to hire more labour it may have to advertise for new workers, but once the new workers are employed the advertisements are no longer necessary. This advertising cost is an adjustment cost: it is incurred solely because the firm wishes to hire more workers, since no advertisement is needed to retain workers already employed. Firms must shop around, search, and collect information just as consumers do. Moreover, changes in input quantities have to be planned and organized over and above the management of ongoing activities. All this absorbs resources and hence imposes costs of adjustment.

If actual input levels differ from cost minimizing levels the firm will gain from changes in input levels. These changes will, however, in themselves involve adjustment costs and so the firm has a problem of finding the *optimal rate of adjustment* by balancing the benefits (reduced production costs) against the adjustment costs of the changes.

Such problems are complicated (though not impossibly so) and we will therefore adopt here the crude simplification of regarding fixed and variable inputs as *polar cases* of adjustment costs. Variable inputs can be thought of as having zero adjustment costs and fixed inputs as having infinite adjustment costs for changes within period 0. The reader should remember that the terms 'long-run' and 'short-run' are based on these polar cases and that the rate of adjustment of inputs by the firm is not solely technologically determined: it depends on an economic decision balancing the benefits and costs of adjustment. We consider adjustment costs again in section 14E.

Opportunity costs

Before we can analyse the firm's cost minimization problems we must define the 'cost' of an input to the firm. The *marginal opportunity cost* of an input is the value of the alternative foregone by the use of an additional unit of that input by the firm. If the additional unit is not already owned or hired by the firm then it must be bought or rented, and the marginal opportunity cost is the market price or rental of the input. If the additional unit used is already owned or rented there is no additional cash outlay by the firm, but, since the unit could have been sold on the market, the market price is the value of the alternative (selling the unit rather than using it) which is foregone.

In the analysis of this chapter we interpret the 'cost' of an input as its marginal opportunity cost and assume that this is measured for variable inputs by the market price of the input. This assumption may not be valid for a number of reasons:

(a) If the market price of the input to the firm varies with the amount purchased then if the price rises (falls) as the firm buys larger quantities of the input the marginal opportunity cost of the input is greater (less) than its market price to the firm. The reason is that the cost to the firm of an extra unit consists of the market price for that unit *plus* the effect of the change in price on the total cost of the units which the firm has already decided to buy.

We leave to Chapter 14 the analysis of this case and assume throughout this chapter that input prices are fixed as far as the firm is concerned.

(b) The firm may face different market prices for the input depending on whether it wishes to buy or sell it. Purchase taxes may cause the buying price to exceed the selling

price. Markets may be costly to use because of the costs of acquiring information, negotiation, etc., so that a seller may receive a net price below that paid by a buyer. These *transactions costs* may also include fees and commissions paid to agents and brokers. The contract under which an input was hired or bought may create a gap between buying and selling prices. For example, a firm may rent warehouse space under a contract which forbids the firm to re-let. The selling price is therefore zero but the purchase price of additional space is the market price. Again, consider a firm which hires labour under a contract which gives workers the right to a month's notice of dismissal, so that their wages are an inescapable cost over this period. The marginal opportunity cost of the input in the short-run decision problem in such cases is the selling price (zero in the two examples above) for quantities less than the amount already owned or rented and the buying price for larger quantities. In the long run (a month in the labour contract example) the marginal opportunity cost is the market price irrespective of the quantity the firm wishes to use.

The marginal opportunity cost of an input depends on the quantity which the firm wishes to use, the quantity which is already owned or contracted for, the costs of using the input market, and the terms of the contract under which inputs are traded. As the last two examples above indicate, it will also depend on the time horizon of the decision for which the cost calculations are required, i.e. on whether the decision is short- or long-run, or whether the input is fixed or variable.

Exercise 8A

- 1.* If the firm can borrow and lend at the interest rate r per annum what is the opportunity cost of using an infinitely durable asset for one year, with and without a secondhand market in the durable asset? How would significant transaction costs (due to the need to dismantle and transport the asset each time it is sold) affect your answer? Suppose the asset had a finite life?

B. Long-run cost minimization

The firm's long-run cost minimization problem is to formulate a *plan* (an input combination) which will minimize the cost of producing a *specified output* during some period sufficiently far into the future for inputs to be considered fully variable. The firm is assumed to be able to buy inputs or sell inputs that it already owns, at a constant positive price, so that the total cost to be minimized is $\sum p_i z_i$. The production function constraining the minimization is assumed to be strictly quasi-concave and twice continuously differentiable. The long-run cost minimization problem is

$$\begin{aligned} \min_{z_1, \dots, z_n} \sum p_i z_i \quad \text{s.t.} \quad & (i) \ f(z_1, \dots, z_n) \geq y \\ & (ii) \ z_i \geq 0 \quad (i = 1, \dots, n) \end{aligned} \quad [\text{B.1}]$$

where y is the required output level.

Figure 8.1 illustrates a two input version of the problem. The lines C^1, C^2, C^3 are *isocost* lines which show the combinations of the two inputs which have the same total cost. The

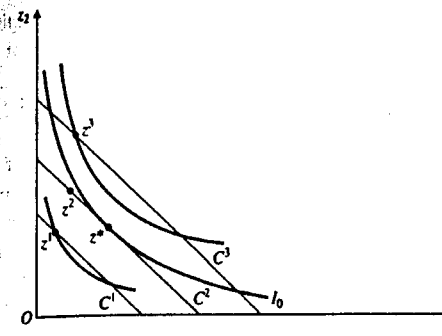


Fig. 8.1

C^1 line, for example, graphs the equation

$$p_1 z_1 + p_2 z_2 = C^1$$

or

$$z_2 = \frac{C^1}{p_2} - \frac{p_1}{p_2} z_1$$

In this case, where the prices of the inputs are independent of the amounts of the inputs bought by the firm, the isocost lines are parallel straight lines with slope

$$\left. \frac{dz_2}{dz_1} \right|_{dC=0} = -\frac{p_1}{p_2} \quad [\text{B.2}]$$

The further from the origin the higher are the total costs represented by the lines: z^2 on C^2 is an input bundle containing more of both inputs than z^1 on C^1 . It must therefore cost more, and since all points on the same isocost line have the same total cost, all points on C^2 cost more than all points on C^1 . I_0 is the isoquant for the required output and, as we argued in section 7A, the solution must be on this isoquant when input prices are positive. The problem is to choose the point on I_0 which has the lowest cost, i.e. is on the lowest isocost line. In this case the least cost input combination is z^* where I_0 is tangent to C^2 . Combinations along lower isocost lines such as C^1 cost less than z^* but do not produce enough output: they are on lower isoquants. Combinations on higher isocost lines such as z^3 on C^3 satisfy the output constraint but have higher costs.

The slope of the isoquant is the negative of the marginal rate of technical substitution between z_1 and z_2 and, in the interior solution illustrated here, cost is minimized where

$$-\frac{p_1}{p_2} = \left. \frac{dz_2}{dz_1} \right|_{y=y^0} = -\text{MRTS}_{21} = -\frac{f_1}{f_2}$$

or

$$\frac{p_1}{p_2} = \frac{f_1}{f_2} \quad [\text{B.3}]$$

The ratio of input prices is equal to the ratio of the marginal products. Rearranging this expression yields

$$\frac{p_1}{f_1} = \frac{p_2}{f_2} \quad [\text{B.4}]$$

as a necessary condition for cost minimization. Now f_1 is the marginal product of z_1 : the rate at which y increases as z_1 increases, and $1/f_1$ is the rate at which z_1 must increase to increase y ; it is approximately the number of units of z_1 required to increase y by one unit. p_1 is the cost of an additional unit of z_1 . p_1 times $1/f_1$ is therefore the cost of increasing the output of y by one unit by increasing the input of z_1 . p_2/f_2 has a similar interpretation. When costs are minimized the firm would be indifferent between increasing y by increasing z_1 or z_2 .

The effect on total cost is the same whichever input is varied so as to increase output by one unit, when inputs are chosen optimally. $p_1/f_1 = p_2/f_2 = LMC$ is therefore the long-run marginal cost of extra output to the firm: the rate at which cost increases as y increases when cost is minimized for every level of y and all inputs are variable.

In section 7A we introduced two distinct but related definitions of efficiency (output efficiency and technical efficiency) and we now introduce a third: *economic efficiency*. An input combination is economically efficient when it minimizes the cost of producing a given output. It is important to be clear about the relationships of these three types of efficiency: economic efficiency implies technical efficiency, which implies output efficiency, but none of the converse implications hold.

Method of Lagrange in the cost minimization problem

Since the solution to [B.1] will satisfy $y = f(z_1, \dots, z_n)$ on our assumptions about input prices and technology, we can, if we also assume that all inputs are used in positive quantities in the solution, analyse the solution to [B.1] by forming the Lagrange function

$$L = \sum p_i z_i + \lambda [y - f(z_1, \dots, z_n)] \quad [\text{B.5}]$$

First-order conditions for a minimum of L are

$$\begin{aligned} \frac{\partial L}{\partial z_i} &= p_i - \lambda f_i = 0 \quad (i = 1, \dots, n) \\ \frac{\partial L}{\partial \lambda} &= y - f(z_1, \dots, z_n) = 0 \end{aligned} \quad [\text{B.6}]$$

and by writing the conditions on z_i as $p_i = \lambda f_i$ and dividing the i th condition by the j th we have the n -input extension of [B.3]:

$$\frac{p_i}{p_j} = \frac{f_i}{f_j} \quad (j = 1, \dots, n, j \neq i) \quad [\text{B.7}]$$

We can, as in all economic problems using Lagrange techniques, give an economic interpretation to λ . Recall section 2F, in which it was shown that the optimal value of λ is the rate at which the optimized value of the objective function increases as the constraint parameter is increased. In [B.1] the objective function is total cost and the constraint

parameter is output, so that the optimal value of λ is the rate at which cost increases as output increases, i.e. long-run marginal cost (LMC) so that

$$\lambda = \frac{\partial C}{\partial y} = LMC$$

where C is the minimized value of $\sum p_i z_i$. This interpretation is supported by writing the conditions [B.6] as

$$\frac{p_1}{f_1} = \dots = \frac{p_n}{f_n} = \lambda = LMC \quad [\text{B.8}]$$

and using the previous discussion of the two-input case in [B.4].

Cost function

The cost-minimizing input levels which solve [B.1] are the *conditional input demands* and are functions of the prices of the inputs and the output level required:

$$z_i^* = z_i(p_1, \dots, p_n, y) = z_i(p, y) \quad [\text{B.9}]$$

The input demands are conditional on the output of the firm so a full explanation of the firm's input demands must include a theory of its choice of output level. The results we derive from cost minimization can be used in any complete model of the firm which requires that the cost of producing the firm's optimal output be minimized.

The *cost function* relates the minimized cost of the firm to input prices and output:

$$C = \sum p_i z_i^* = pz(p, y) = C(p, y) \quad [\text{B.10}]$$

We are interested in the effects of changes in input prices and output on the firm's conditional input demands and on its minimized cost. From [B.10] properties of $z(p, y)$ and $C(p, y)$ are clearly related.

The reader will have noticed that the firm's problem of minimizing the cost of producing a specified output y is remarkably similar in form to the consumer's problem in section 4A of minimizing the expenditure necessary to achieve a particular utility level. Indeed, if z denoted a consumption bundle, y utility, $f(z)$ the utility function and p the price vector of consumption goods, [B.1] would be identical to the consumer's expenditure minimization problem. This means that the results we derived concerning the expenditure minimization problem carry over directly to the firm's cost minimization problem. All that is required is a suitable relabelling so that instead of the Hicksian, constant utility demands $h_i(p, u)$ for goods by the consumer we refer to conditional input demands $z_i(p, y)$ and instead of the expenditure function $m(p, u)$ we refer to the firm's cost function $C(p, y)$.

In section 4A we examined the properties of the expenditure function. We can restate some of them here in terms of the firm's cost function:

- (a) $C(p, y)$ is increasing in y and non-decreasing in p ;
- (b) $C(p, y)$ is linear homogeneous in p : $C(kp, y) = kC(p, y)$;

- (c) $C(p, y)$ is continuous and concave in p ;
 (d) *Shephard's lemma*: $\partial C(p, y)/\partial p_i = z_i(p, y)$.

We make extensive use of these properties in our analysis of the effects of p and y on the cost function and the conditional input demands.

Since we have already derived the properties in section 4A we leave it to the reader to apply the arguments in that section (with suitable relabelling) to the firm's cost function. We will, however, present an alternative proof of Shephard's lemma which is neater, though perhaps less intuitive, than the one given in section 4A. Consider the function

$$G(p, p^0, y) = C(p, y) - pz(p^0, y) \leq 0 \quad [\text{B.11}]$$

This expression cannot be positive because $z(p^0, y)$ is the cost minimizing input bundle at input prices p^0 and it cannot yield a smaller cost of producing y at some other price vector p than the bundle $z(p, y)$ which is cost minimizing at p . However, at $p = p^0$, $z(p^0, y)$ is optimal, the cost function is $C(p^0, y) = p^0 z(p^0, y)$ and $G(p^0, p^0, y) = 0$. Thus $G(p, p^0, y)$ is maximized with respect to p at $p = p^0$. Hence at $p = p^0$ the partial derivatives of G with respect to p_i must be equal to zero:

$$\left. \frac{\partial G(p, p^0, y)}{\partial p_i} \right|_{p=p^0} = \left. \frac{\partial C(p, y)}{\partial p_i} \right|_{p=p^0} - z_i(p^0, y) = 0$$

Since $C_i(p^0, y) = z_i(p^0, y)$ must be true for all p^0 we have established Shephard's lemma.

The cost function is useful because it contains all the economically relevant information about the firm's technology. If we know the cost function we can discover the cost minimizing input bundle $z(p, y)$ for any output y at any prices p by using Shephard's lemma. Thus we can find a set of input combinations which can be used to produce y and since we know that cost minimization implies that the firm is output efficient this set must be a subset of the isoquant for y . There may be other input combinations which are also on the isoquant for y but because they are not cost minimizing at *any* p they are not economically relevant: no cost minimizing firm would ever choose them. Fig. 8.2 shows

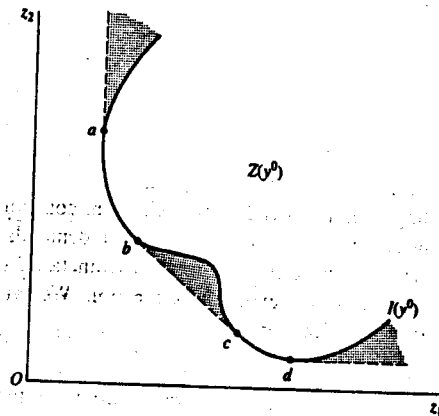


Fig. 8.2

a rather bizarre technology in which the input requirement set $Z(y^0)$ is non-convex and the non-economic region is not empty (notice the positive sloped segments of the isoquant $I(y^0)$). The isoquant $I(y^0)$ is the whole of the lower boundary of $Z(y^0)$. As the reader should check by drawing negatively sloped isocost lines, no cost-minimizing firm facing positive input prices will ever choose to produce y^0 from an input combination in the segments of $I(y^0)$ 'north-east' of a or d or between b and c . Such points on the isoquant are feasible but will never be chosen by a cost-minimizing firm which wished to produce y^0 . The only input choices which can ever be observed are those between a and b and between c and d . These are identical with the input choices made by a cost minimizing firm which faces a technology giving rise to an input requirement set $Z^*(y^0)$ consisting of the $Z(y^0)$ plus the shaded areas. Thus, although knowledge of the cost function does not tell us everything about the technology, it does convey all the information which is relevant for modelling cost minimizing firms. Note that $Z^*(y^0)$ is convex even though $Z(y^0)$ is not, so that there is no loss in generality in assuming that cost-minimizing firms face quasi-concave production functions.

We assumed that $f(z)$ was strictly quasi-concave and twice continuously differentiable, because, as we noted in Chapter 2, these assumptions enable us to use calculus methods in studying optimization problems. We can summarize the economically relevant features of the technology in the cost function and the cost function has the properties listed above under much weaker conditions on the technology than are required to use Lagrangean methods to directly analyse cost minimizing input choices.

Input choice and output level

Figure 8.3 illustrates the effects of changes in y on the optimal cost minimizing input choices. z^0, z^1, z^2 are the input choices for producing output levels y^0, y^1, y^2 at minimum cost of C^0, C^1, C^2 respectively. The *expansion path* EP is the locus of optimal input combinations traced out as the required output varies with input prices held constant. Here EP is positively sloped indicating that increases in y cause increases in both inputs. However, with a different technology the expansion path can be negatively sloped over

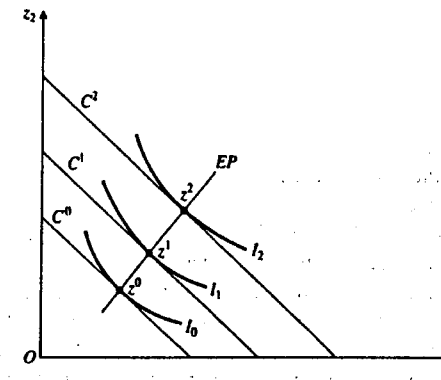


Fig. 8.3

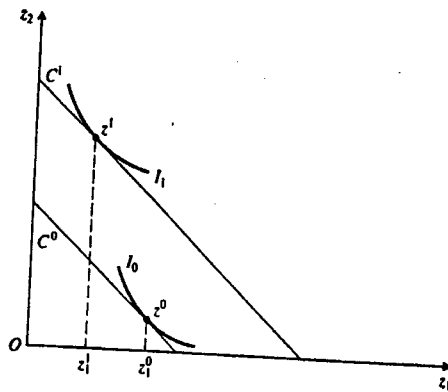


Fig. 8.4

part of its range, as in Fig. 8.4. Here as y increases from y^0 to y^1 the amount of z_1 used declines from z_1^0 to z_1^1 . Over this range z_1 is an *inferior or regressive* input and z_2 is *normal*. (Why must at least one input be normal?)

In section 7B we showed that if the production function is homothetic then the slopes of isoquants are constant along rays from the origin. Since [B.7] is necessary for cost minimization, we see that *if the production function is homothetic input proportions are the same at all output levels*, and the expansion path will be a ray from the origin. Only changes in relative input prices cause changes in input proportions.

Long-run cost curves

The relationship between long-run cost and the level of output can be read off from the expansion path in Fig. 8.3 and graphed in Fig. 8.5(a). The isocost lines give total cost and the isoquants the output level for each point on EP . For example, the (minimized) cost of y^0 is C^0 , of y^1 is C^1 and of y^2 is C^2 . In Fig. 8.5(a) these outputs are plotted along the horizontal axis and the corresponding total costs along the vertical axis. LTC is the long-run total cost curve derived from minimizing cost for each level of output when all inputs are variable. As drawn, it embodies some particular assumptions about technology which will shortly be clarified.

The long-run average and marginal cost curves (LAC and LMC) which are plotted in part (b) of Fig. 8.5 are derived in turn from the LTC curve. The long-run average cost of producing y^0 is C^0/y^0 and this is the slope of the line OA in (a), which goes from the origin to the point on the LTC curve where $y = y^0$ and $C = C^0$. The LAC curve plots the slopes of the rays from the origin to the LTC curve. The fact that the rays get steadily flatter up to point B , and then steeper, accounts for the U-shaped LAC curve.

Since long-run marginal cost is the rate at which long-run cost increases as output increases ($LMC = \partial C / \partial y$) the LMC curve is derived by plotting the slope of the LTC curve from below at the point where LAC is at a minimum, since at output y^2 the ray from the origin OB is also tangent to the curve. It can be shown that this relationship must always hold by the same reasoning as was applied to the relationship between average

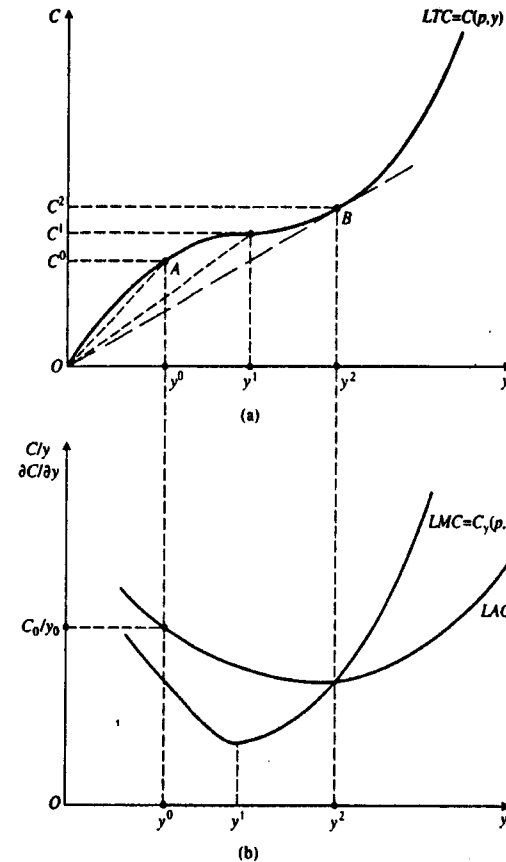


Fig. 8.5

and marginal product curves in section 7C. (See Question 1, Exercise 8B.) Note also that the output y^1 at which LMC is a minimum is the point of inflexion of the LTC curve, and that LAC is decreasing through this point (the rays in (a) are still getting flatter). Again the curvature of the LTC curve in (a), with its slope, though always positive, at first falling and then rising, implies the U-shaped LMC curve in (b).

Economies of scale and returns to scale

The *elasticity of cost with respect to output* is a measure of the responsiveness of cost to output changes. It is defined as the proportionate change in cost divided by the proportionate change in output:

$$E_y = C_y(p, y) \frac{y}{C(p, y)} = \frac{LMC}{LAC} \quad [B.12]$$

(remember that $C_y = LMC$ and $C/y = LAC$). The cost function has *economies of scale* if $LMC/LAC < 1$ and *diseconomies of scale* if $LMC/LAC > 1$. Since $LMC < LAC$ implies that LAC is decreasing with y , there are economies of scale when the LAC curve is falling. Conversely, there are diseconomies when the LAC curve is rising. In Fig. 8.5 there are economies of scale up to y^2 and diseconomies thereafter.

The relationship between output and costs depends on the underlying technology. Suppose that there are increasing returns to scale and $z(p, y^0)$ is cost minimizing for output y^0 at prices p . It will be possible to produce an output twice as large as y^0 from $sz(p, y^0)$ (defined by $f(s(z(p, y^0))) = 2y^0$) where $s < 2$. Hence cost will be less than double when output doubles and so there are economies of scale. (Note that if $sz(p, y^0)$ is not cost minimizing for $y = 2y^0$ the argument holds *a fortiori*.) Now suppose that there are decreasing returns to scale and $z(p, y^0)$ is cost minimizing for output y^0 at prices p . It will be possible to produce an output half as large as y^0 from $sz(p, y^0)$ (defined by $f(s(z(p, y^0))) = \frac{1}{2}y^0$) where $s < \frac{1}{2}$. Hence cost will be more than halved when output is halved and so there are diseconomies of scale. Thus we have established

$$E_y^c = \frac{LMC}{LAC} \leq 1 \Leftrightarrow E \geq 1 \quad [\text{B.13}]$$

where E is the elasticity of output with respect to scale.

Homotheticity and the cost function*

Recall that a homothetic production function can be written in the form $g(z) = F(f(z))$ where $F' > 0$ and $f(z)$ is linear homogeneous. With a homothetic production function the cost minimizing input proportions are independent of the output required so that if $z(p, y^0)$ produces y^0 at minimum cost, $s(y)z(p, y^0)$ will produce y at minimum cost. $s(y)$ is the proportionate change in inputs required to produce y and so $C(p, y) = s(y)C(p, y^0)$. But $F(s(y)f(z(p, y^0))) = y$ implies that

$$s(y)f(z(p, y^0)) = F^{-1}(y) = a(y)$$

where $a(y) = F^{-1}(y)$ is the inverse of $F(\cdot)$. Hence the cost of producing y is

$$C(p, y) = s(y)C(p, y^0) = a(y)C(p, y^0)/f(z(p, y^0)) = a(y)b(p) \quad [\text{B.14}]$$

where $b(p) = C(p, y^0)/f(z(p, y^0))$. (Compare the consumer's expenditure function in the case of homothetic preferences.) Thus if the production function is homothetic then the cost function can be written in the form $C(p, y) = a(y)b(p)$.

Homogeneous functions are homothetic so the reader can check that when the production function is homogeneous of degree n , $a(y)$ has the form $y^{1/n}$. In particular, if the production function is linear homogeneous, cost is directly proportional to output since a proportional increase in output requires the same proportional increase in inputs. The reader is asked to show (see Question 5)

$$E_y^c = 1/E \quad [\text{B.15}]$$

i.e. the elasticity of cost with respect to output is the reciprocal of the scale elasticity if the production function is homothetic. Since the cost minimizing input proportions do not vary

with output if the production function is homothetic, changes in output require only changes in scale. Hence the relationship between cost and output depends only on the relationship between output and scale. Cost varies proportionately with scale but output may vary proportionately more or less than scale. For example, with increasing returns ($E > 1$) costs will vary less than proportionately with output and there will be economies of scale ($E_y^c < 1$).

Input prices and conditional input demands

We can use the properties of the cost function to examine the relationship between the prices of inputs and the conditional input demands. The partial derivative of $C(p, y)$ with respect to p_i is $z_i(p, y)$ (property (d) – Shephard's Lemma). Since the cost function is homogeneous of degree one in p (property (b)) the partial derivative of $C(p, y)$ with respect to p_i is homogeneous of degree zero (recall the discussion of homogeneous functions in section 7B). Hence $z_i(p, y)$ is homogeneous of degree zero and equal proportionate changes in all input prices have no effect on the cost minimizing input choices: $z(sp, y) = z(p, y)$. If all input prices change in the same proportion the slopes of the isocost lines in Fig. 8.1 are unchanged and thus the z^* where the isocost line is tangent to the isoquant is also unchanged. Less informally (and not requiring any smoothness restrictions on technology): if $pz^* \leq pz$ for all z in $Z(y)$, so that z^* is cost minimizing at p , then $spz^* \leq spz$ for all z in $Z(y)$ and z^* is also cost minimizing at prices sp .

Next, suppose that p changes from p^0 to p^1 where p^1 is not necessarily proportional to p^0 . The cost-minimizing input choice at p satisfies $pz(p, y) \leq pz$ for all z in $Z(y)$ and thus $z(p^1, y)$ cannot cost less at p^0 than $z(p^0, y)$:

$$p^0 z(p^0, y) - p^0 z(p^1, y) = p^0 [z(p^0, y) - z(p^1, y)] \leq 0 \quad [\text{B.16}]$$

Similarly, $z(p^0, y)$ cannot cost less at p^1 than $z(p^1, y)$:

$$p^1 z(p^0, y) - p^1 z(p^1, y) = p^1 [z(p^0, y) - z(p^1, y)] \geq 0 \quad [\text{B.17}]$$

Subtracting [B.17] from [B.16] gives

$$(p^0 - p^1)[z(p^0, y) - z(p^1, y)] \leq 0 \quad [\text{B.18}]$$

so that the sum of the price changes times the input demand changes cannot be positive.

If only p_1 changes the resulting change in $z_1(p, y)$ is the *own price input substitution effect*. [B.18] reduces to $(p_1^0 - p_1^1)[z_1(p^0, y) - z_1(p^1, y)] \leq 0$ and we can see that the *own price input substitution effect is non-positive*. Figure 8.6 illustrates for the case in which the isoquants are smooth. With input prices initially giving rise to isocost lines like C^0 the cost minimizing input choice is z^0 , where the isoquant I_0 is tangent to C^0 . Let the price of z_1 increase. The isocost lines will pivot about their intercepts on the z_2 axis and will become steeper. The new optimal choice is z^1 where the isocost line C^1 is tangent to I_0 . The increase in the relative price of input 1 must be to reduce the conditional demand for it steepens the isocost lines and the isoquants become steeper as z_1 is reduced (z_2 is substituted for z_1).

We can reach the same conclusion by using Shephard's lemma and the concavity of the cost function (property (c)). Concavity places restrictions on the second-order partial derivatives of the cost function. In particular, the second-order own partials $C_{ii}(p, y)$ must

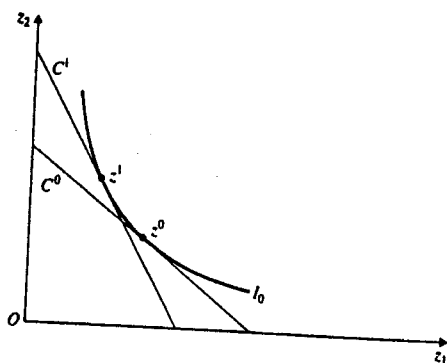


Fig. 8.6

be non-positive, which, using the fact that $C_i(p, y) = z_i(p, y)$ (Shephard's lemma), implies that

$$\frac{\partial z_i(p, y)}{\partial p_i} = \frac{\partial C_i(p, y)}{\partial p_i} \leq 0 \quad [\text{B.19}]$$

In section 7B we introduced the elasticity of substitution as a measure of the relationship between the slope of the isoquant and the input ratio z_1/z_2 and indicated that the concept is useful in analysing the demand for inputs. We must defer a full discussion to Chapter 14 since we do not yet have any model of the firm's choice of output but note from Fig. 8.6 that the effect of changes in relative input prices p_1/p_2 on the input proportions used to produce a given output depends on the curvature of the isoquant. Cost minimization requires that the slope of the isoquant be equal to the slope of the isocost line $-p_1/p_2$ and so the greater is the elasticity of substitution the greater will be the input substitution effects of changes in input prices.

Effects of input price changes on costs

The effect of a proportionate increase in p on the firm's total and average cost is straightforward since we know that $C(p, y)$ will change in the same proportion (property (b) – linear homogeneity) and thus so will $C(p, y)/y$. Thus the firm's total and average cost curves will shift upward by the same proportion. Since long-run marginal cost is p_i/f_i (see [B.8]) its LMC curve will also be shifted up proportionately.

The effects on LTC and LAC of a change in the price of one input only are also fairly straightforward. Using Shephard's lemma, the elasticity of cost with respect to p_i is

$$E_{p_i}^C = \frac{\partial C(p, y)}{\partial p_i} \cdot \frac{p_i}{C(p, y)} = \frac{z_i(p, y)p_i}{C(p, y)} \quad [\text{B.20}]$$

The responsiveness of cost to a change in the price of a single input is equal to the proportion of total cost accounted for by expenditure on that input. Since average cost is $C(p, y)/y$ and y is held constant in determining the effect of p_i on average cost, we leave

it to the reader to establish that the elasticity of average cost with respect to p_i is also equal to the expenditure share of input i .

The effect of a given rise in p_i on the LTC and LAC curves will be to shift the LTC and LAC curves upward vertically by an amount dependent on the proportion of total cost which is spent on z_i . This does not mean that the curves shift by the same proportion for all output levels since the proportion of C spent on z_i may well vary with the output level. The effect of the change in p_i may be to increase or lower the output level at which LAC is a minimum and to increase or decrease the slope of the LAC curve at any output level. The precise effects will depend on the production function. For example, if it has linear expansion paths (MRTS constant along rays from the origin) then the proportion of total cost spent on the i th input will be constant since input proportions are constant along all expansion paths. Hence the LTC and LAC curves will shift vertically upward in the same proportion for all output levels and the output at which LAC is at a minimum will be unchanged.

The effect on the firm's LMC curve is less easy to predict without knowledge of the production function. The reason for this can be shown in Fig. 8.7. The initial input prices give rise to isocost lines C_0, C_2 and optimal input bundles z^0, z^2 for outputs of y^0 and y^1 . The new higher price of p_1 gives isocost lines C_1, C_3 and optimal bundles z^1, z^3 for outputs of y^0 and y^1 . The change in total cost for the change in output $\Delta y = y^1 - y^0$ with the initial lower price of z_1 is $\Delta C = C_2 - C_0$ and this can be measured in the diagram by p_2 times the distance AB . Similarly, with the higher price of z_1 the change in cost caused by a change in output from y^0 to y^1 is $\Delta C' = C_3 - C_1$ and is measured by p_2 times the distance DC . In Fig. 8.7 $\Delta C' > \Delta C$ and thus the effect of the rise in p_1 is to increase the marginal cost of Δy . However, with a differently shaped isoquant it is possible that $\Delta C' < \Delta C$. (Draw the diagram.) Hence it is impossible to predict the effect of a rise in p_i on marginal cost without knowledge of the production function.

Use of Shephard's lemma shows us exactly what is required for marginal cost $C_y(p, y)$ to increase or decrease with p_i . Since cross-partial derivatives do not depend on the order of differentiation

$$\frac{\partial^2 C(p, y)}{\partial y \partial p_i} = \frac{\partial^2 C(p, y)}{\partial p_i \partial y} = \frac{\partial z_i(p, y)}{\partial y} \quad [\text{B.21}]$$

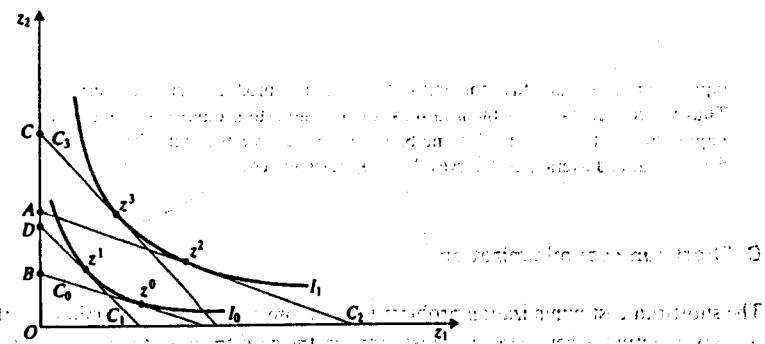


Fig. 8.7

Now $\partial z_i(p, y)/\partial y$ is the effect on the demand for input i of an increase in output with input prices held constant and is positive or negative as z_i is a normal or regressive input. Hence an increase in p_i increases marginal cost if and only if z_i is a normal input.

If the expansion path is a ray from the origin, so that all inputs are normal, marginal cost must increase with p_i . Thus [B.21] implies that if the production function is homothetic marginal cost increases with p_i .

Exercise 8B

- Fixed proportions technology.** Illustrate the solution to the cost-minimization problem if the firm has the fixed proportions Leontief technology $y = \min(z_1/\beta_1, z_2/\beta_2)$. Show that the cost function is $C(p, y) = y(\beta_1 p_1 + \beta_2 p_2)$. Sketch the cost curves. Derive the conditional input demand functions.
- Linear technology.** Suppose the production function is $y = \alpha_1 z_1 + \alpha_2 z_2$. Sketch the firm's isoquants and the solution to its cost minimization problem. Show that the cost function is $C(p, y) = y \min(p_1/\alpha_1, p_2/\alpha_2)$. Sketch the cost curves. Derive the conditional input demand functions. Compare the results to those in question 1.
- Cobb-Douglas technology.** Show that the cost function for a firm with the constant returns Cobb-Douglas production function $y = Az_1^\alpha z_2^{1-\alpha}$ of Question 5, Exercise 7A is $C(p, y) = y p_1^\alpha p_2^{1-\alpha} B$ where B is a function of A and α only. Sketch the cost curves. Derive the conditional input demands.
- Assume that the firm owns z_1^0 units of z_1 and that the constant buying and selling prices of z_1 differ because of transaction costs. Draw the firm's isocost lines and sketch the solution to its cost-minimization problem. Show the expansion path and draw the long-run cost curves.
- Homotheticity and the cost function.** (a) Show that if the production function is homogeneous of degree n then the cost function can be written as $C(p, y) = y^{1/n} b(p)$. (b) Show that $E_y^C = 1/E$ if the production function is homothetic.
- Elasticity of substitution.** What is the relationship between the elasticity of substitution and the effect of a change in relative input prices on the firm's relative expenditure ($p_1 z_1 / p_2 z_2$) on its inputs in the case of a two input production function?
- Indivisibility and the cost function.** Suppose that the firm uses a single indivisible input to produce y and that one unit of the input can produce \bar{y} units of output. Thus to produce $0 < y \leq \bar{y}$ the firm must use one unit, to produce $\bar{y} < y \leq 2\bar{y}$ would require two units and so on. Assume that the input costs p per unit. Sketch the firm's total and average cost curves. What is marginal cost?

C. Short-run cost minimization

The short-run cost minimization problem is to choose a (z_1, z_2) pair to minimize the cost of a given output, when there are constraints on the adjustment of the fixed input z_2 . The short-run cost function and associated curves show the relationship between y and

minimized cost and are derived from the minimization problem. The constraints on z_2 , and hence the short-run cost function, may take a variety of forms (see section A). We will assume that the constraint is of the form $z_2 \leq z_2^0$. There is a fixed ceiling on the amount of z_2 available in the period but, since inputs are assumed divisible, the firm can choose to use less if it wants to. To bring out the circumstances under which it would or would not choose to, we consider the following two cases:

(a) The firm faces a quota or ration on z_2 and pays the market price p_2 for units of z_2 bought, up to a maximum of z_2^0 units. The marginal opportunity cost of z_2 is p_2 for $z_2 \leq z_2^0$ and effectively infinite for $z_2 > z_2^0$. Short-run total cost is $p_1 z_1 + p_2 z_2$ and the short-run isocost lines have a slope (the negative of the ratio of marginal opportunity costs) of $-p_1/p_2$ for $z_2 < z_2^0$. An example of this case would be where the firm has a leasing agreement under which it may lease units of z_2 up to some stipulated maximum per period, and it only pays for what it uses. Since inputs are assumed divisible, this implies that it is free to use and pay for less z_2 than the maximum z_2^0 .

(b) The firm has contracted to pay $p_2 z_2^0$ for the fixed input regardless of whether it uses less of it than z_2^0 or not. Equivalently, the firm may own z_2^0 units of z_2 and transactions costs or the absence of a market prevent the firm from selling those units of z_2 it does not want to use. Hence, unlike case (a), the existence of a fixed input creates a fixed cost. This is the essence of the difference between cases (a) and (b), and reflects the fact that a 'fixed input' i.e. one which is subject to a maximum level of use, need not imply a fixed cost – it all depends on the nature of the relevant contract into which the firm has entered. Here, the short-run total cost is $p_1 z_1 + p_2 z_2^0$, where $p_1 z_1$ is total variable cost and $p_2 z_2^0$ is total fixed cost. Since changes in z_2 below the capacity level z_2^0 cause no change in costs, the marginal opportunity cost of z_2 is zero for $z_2 < z_2^0$, and is effectively infinite for $z_2 > z_2^0$ (no more can be had at any price).

The derivation of the short-run cost curves for cases (a) and (b), and their relation to the long-run cost curve, are shown in Figs 8.8 and 8.9. In Fig. 8.8, the curve EP again represents the expansion path – the locus of points of tangency of price lines of slope $-p_1/p_2$ with isoquants. The cost/output pairs lying along EP are then plotted as the long-run total cost curve in Fig. 8.9. In the figures we show just three such points. Output

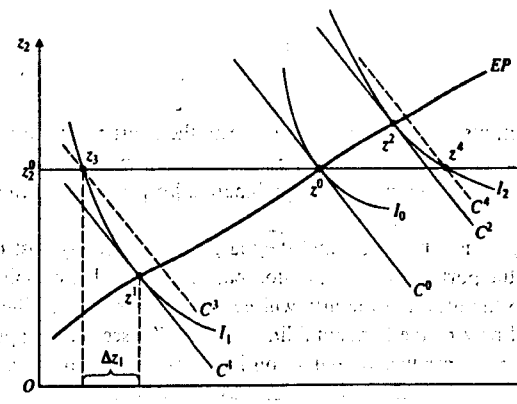


Fig. 8.8

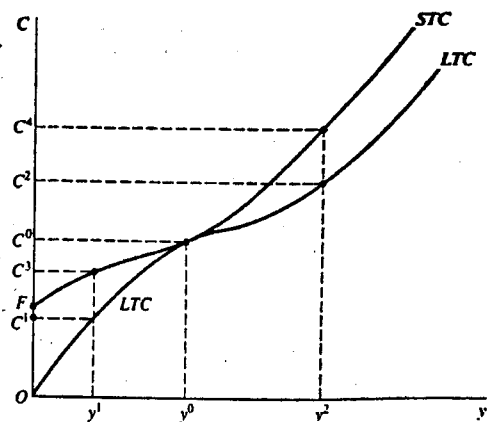


Fig. 8.9

y^1 , corresponding to isoquant I_1 , and the associated minimized cost C^1 , output y^0 , corresponding to isoquant I_0 , and its minimized cost C^0 , and output y^2 with cost C^2 . We now consider the analysis for the short run.

Take first case (a). For $z_2 < z_2^0$, the marginal opportunity cost of z_2 is identical to that in the long run. For example, if the firm wished to produce output y^1 then the solution to its cost-minimizing problem is point z^1 in the figure (supply the details of the argument). Thus at such an output the firm would choose to use less than z_2^0 , the maximum available. A similar result holds for all outputs up to and including y^0 , corresponding to isoquant I^0 . (Again, supply the argument.) Thus for case (a) the expansion path *coincides with* EP up to and including the point z^0 and over the corresponding range of outputs the short-run total cost curve coincides with the long-run total cost curve in this case.

For outputs greater than y^0 , to move further along EP would require amounts of $z_2 > z_2^0$, which are unavailable to the firm. For example, output y^2 corresponding to isoquant I_2 would require an amount of z_2 which is the coordinate of point z^2 in the figure. To produce y^2 , the best the firm can do is to choose point z^4 , using the fixed input to capacity at z_2^0 , and a greater amount of the variable input z_1 , than at z^2 .

It follows that at such an output the total production cost in the short run will be greater than in the long run. Point z^4 lies on the iso-cost line C^4 indicated in the diagram, and $C^4 > C^2$. Hence, for all outputs greater than y_0 in Fig. 8.9, the short-run total cost lies above the long-run total cost. The capacity constraint on z_2 is binding and causes a departure in the short run from the optimal input combination for producing each output level.

In case (b), recall that $p_2 z_2^0$ is a fixed cost and the marginal opportunity cost of z_2 is zero. Since z_2 is divisible, the portion of the expansion path EP to the left of point z^0 in Fig. 8.8 is still available to the firm, but the firm will *not choose* to be on it. The firm's *chosen* expansion path will now be the horizontal line $z^0 z^3 z^0 z^4$. To see this, suppose the firm were to choose point z^1 to produce output y^1 on isoquant I_1 . By moving along I_1 to z^3 , it reduces the amount of z_1 by Δz_1 and therefore saves costs equal to $p_1 \Delta z_1$. There is no corresponding increase in cost due to the increased use of z_2 because its marginal

Opportunity cost is zero: all costs associated with z_2 are fixed and do not vary with the level of use. Hence it always pays the firm to use z_2 to capacity even when it has the (technological) option of not doing so.

This argument can be repeated at all outputs up to y^0 . For outputs above y^0 the earlier argument again holds – no more than z_2^0 can be used to produce any such output. Thus in case (b) the entire short-run expansion path is the horizontal line through z_2^0 . (This conclusion may have to be qualified where this line intersects a ridge line. See Question 3, Exercise 8C.)

The implications of this for the *STC* curve in case (b) are easy to see. At all outputs below y^0 total costs, though minimized *given the capacity constraint*, are higher than in the long run. At a zero output the fixed cost $p_2 z_2^0$ must still be paid, and the intercept *OF* of the *STC* curve in Fig. 8.9 represents this. As output increases *STC* lies above *LTC* (compare C^3 , the cost of input combination z^3 , with C^1 in Fig. 8.8) but converges to it. At y^0 long-run and short-run costs are equal. This is because y^0 is the unique output level with the property that the fixed input level z_2^0 is actually the *optimal* long-run z_2 -level for the output. For outputs above y^0 input combinations are again sub-optimal in the short run, *STC* lies above *LTC* and diverges steadily from it.

Thus we conclude that in case (a) given the input constraint $z_2 \leq z_2^0$, the short-run total cost curve coincides with the long-run total cost curve up to output y^0 (the unique output for which z_2^0 is in fact optimal) and then is the *STC* curve shown in Fig. 8.9. In case (b), on the other hand, the short-run total cost curve is the entire *STC* curve.

Short-run average and marginal cost

We can now derive the short-run average and marginal cost curves from Fig. 8.9 for case (b), leaving the simpler case (a) (in which there are no fixed costs) to the reader. The short-run average and marginal curves are derived in the same way as for the long-run curves in section B and are shown in Fig. 8.10 together with the long-run curves. *SAC* is the short-run average cost, *SMC* the short-run marginal cost curve. Notice that *SMC* cuts *SAC* from below at the output at which *SAC* is at minimum. *SAC* lies above *LAC* for

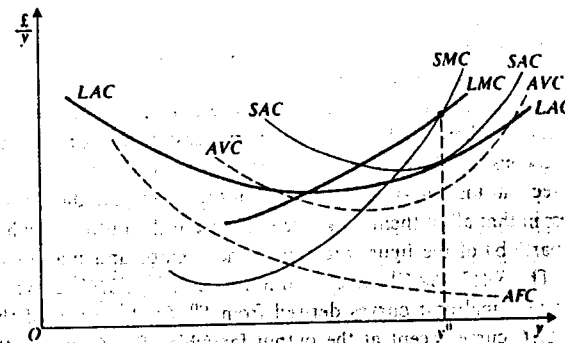


Fig. 8.10

outputs other than y^0 since short-run total cost exceeds long-run total cost for output other than y^0 . Letting $S(y)$ be short-run cost we have $S(y) \geq C(y)$ and hence $S(y)/y \geq C(y)/y$. Short-run average cost is never less than long-run average cost. SAC is tangent to LAC at y^0 because $S(y)$ is tangent to $C(y)$ at y^0 . Differentiating $SAC = S(y)/y$ with respect to y gives

$$\frac{1}{(y)^2} \left[\frac{dS}{dy} \cdot y - S \right]$$

but at y^0 , dS/dy equals dC/dy and $S = C$, so that the slope of SAC equals the slope of LAC . Note also that the tangency of S and C at y^0 implies that SMC equals LMC at y^0 since short- and long-run marginal costs are the slopes of the short-run and long-run total cost curves respectively.

In case (b) short-run cost is the sum of variable cost (VC) and fixed cost (FC):

$$S = VC + FC = p_1 z_1 + p_2 z_2^0$$

where z_1 varies with y . In Fig. 8.10 the dashed AVC curve plots average variable cost $p_1 z_1/y$ and the AFC curve average fixed cost ($p_2 z_2^0/y$) which is a rectangular hyperbola. y/z_1 is the average product AP_1 of z_1 (see section 7C) and so

$$AVC = \frac{p_1 z_1}{y} = \frac{p_1}{AP_1}$$

By similar arguments to those used in the long-run case

$$SMC = \frac{p_1}{f_1} = \frac{p_1}{MP_1}$$

The reader should compare the relationship between the short-run average and marginal cost curves shown in Fig. 8.10 with that between the average and marginal product curves of Chapter 7, Fig. 7.6. The general shapes of the former are the inverse of those of the latter, because of [C.2] and [C.3].

The envelope property

Fixing the z_2 constraint at different levels will generate different short-run cost curves, each of which, in case (b), will lie above the long-run curve except where they are tangent to it at the output for which the constrained level of z_2 is the long-run cost minimizing level. If the expansion path is upward sloping as in Fig. 8.8 the short-run and long-run cost curves will touch at higher levels of output as the fixed level of z_2 is increased. This is illustrated in (a) of Fig. 8.11 where S^0, S^1, S^2 are short-run cost curves for z_2 constraints of $z_2^0 > z_2^1 > z_2^2$. As the constrained level of z_2 varies, more short-run cost curves are generated and we can see that the long-run cost curve C is the lower boundary or envelope of the short-run curves, in that all of them lie above C except at the output at which they are tangent to it. In part (b) of the figure are shown the average and marginal curves derived from part (a). The SAC^0, SAC^1, SAC^2 and SMC^0, SMC^1, SMC^2 curves are the short-run average and marginal cost curves derived from S^0, S^1, S^2 . Each of the SAC curves lies above the LAC curve except at the output for which $S = C$, where they are tangent to it. Hence the LAC curve is the envelope of the SAC curves. The SMC curves,

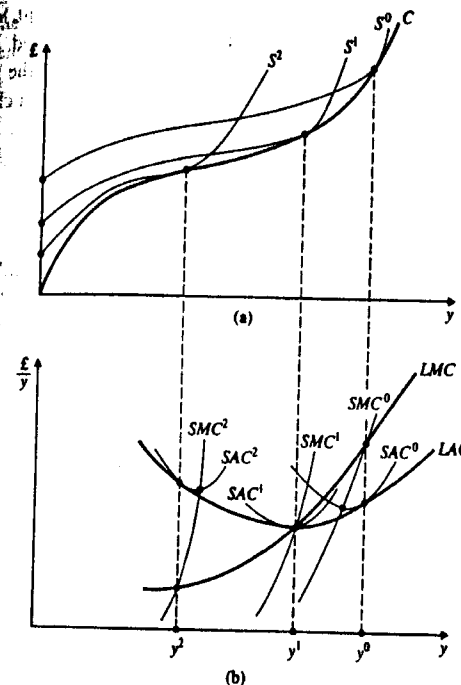


Fig. 8.11

however, cut the LMC curve at the output for which their respective SAC curves are tangent to LAC , and so the LMC curve is not the envelope of the SMC curves. Short-run marginal cost may be greater or less than long-run depending on the output and the level of the fixed input. When the fixed input is at the long-run cost minimizing level for a particular output level SMC equals LMC . In the neighbourhood of this point for larger outputs SMC will exceed LMC , indicating that it will cost more in the short run to expand output than in the long run. On the other hand, at smaller outputs than that for which the fixed z_2 is optimal, short-run marginal costs are below long-run marginal costs. This is because output expansion over this range is improving the rate of utilization of the fixed input – the short-run input combinations are converging toward the long-run input combination (cf. Fig. 8.8.).

This relationship between SMC and LMC is derived from that between the STC and LTC curves in the neighbourhood of the output level at which the fixed input is at its optimal long-run level. Since the STC curve is tangent to the LTC curve from above at y^0 the slope of the STC curve (SMC) must be less than that of the LTC curve (LMC) for $y < y^0$ and greater for $y > y^0$ for some neighbourhood of y^0 :

$$\left. \frac{\partial SMC}{\partial y} \right|_{y=y^0} > \left. \frac{\partial LMC}{\partial y} \right|_{y=y^0} \quad [C.4]$$

However, it is possible to construct *LTC* curves with the envelope property but having $SMC > (<) LMC$ for some $y < (>) y^0$ outside the immediate neighbourhood of y^0 . (Show this.) The implications of the relationship between *SMC* and *LMC* for the firm's response to output price changes in the short and long runs is examined in the next chapter in Question 4, Exercise 9B.

Comparative statics in the short-run

We have already considered the effect of variations in output on short-run cost and input use in deriving the firm's short-run cost curves. Let us now briefly examine the effect of changes in the price of the variable input on the firm's cost curves. In case (a) defined above the firm's short-run expansion path is its long-run expansion path up to $y = y^0$ and $z_2 = z_2^0$ line thereafter. Hence changes in p_1 will cause the expansion path for $y \leq y^0$ to alter in the same way as the long-run path and so all the remarks relevant to the long-run case apply. For $y > y^0$ the expansion path is identical to the case (b) path, to which we now turn.

In case (b) the expansion path is the $z_2 = z_2^0$ line for all outputs. This path is the same for all levels of p_1 so that the optimal short-run input combination is independent of p_1 . Variable cost is $p_1 z_1$ and average variable cost is $p_1 z_1 / y$, so a given percentage change in p_1 will shift the *VC* and *AVC* curves upward in the same proportion. Since the optimal input bundles do not change when p_1 alters $MP_1 = f_1(z_1, z_2)$ will also be unaffected and so $SMC = p_1 / f_1$ will vary proportionately with p_1 . Compare the analogous results for the long run where the effect of changes in p_1 on *LMC* could not be predicted without detailed knowledge of the production function.

Formal analysis*

The results derived graphically for the case in which there is one variable and one fixed input also hold when there are more than two inputs. Denote the n vector of variable inputs $z_v = (z_{v1}, \dots, z_{vn})$ and let p_v be the corresponding n vector of the prices of the variable inputs. Let the m vector of fixed inputs be $z_k = (z_{k1}, \dots, z_{km})$ and p_k be the corresponding m vector of the prices of the fixed inputs. The firm has contracted to pay for z_{kj}^0 units of the j th fixed input but can use less than this if it wishes, i.e. we consider only case (b) here. z_k^0 is the m vector of constraints on the fixed inputs. The firm's short-run cost minimization problem is

$$\min_{z_v, z_k} p_v z_v + p_k z_k^0 \quad \text{s.t. } y = f(z_v, z_k)$$

$$z_{kj} \geq z_{kj}^0 \quad (j = 1, \dots, m)$$

$$z_{vi} \geq 0 \quad (i = 1, \dots, n)$$

[C.5]

The Lagrangean for the problem is

$$L = p_v z_v + p_k z_k^0 + \lambda [y - f(z_v, z_k)] + \sum_j \mu_j (z_{kj} - z_{kj}^0)$$

[C.6]

Assume that the production function is strictly quasi-concave and twice continuously differentiable and that at the solution to the problem all inputs are used. Then the following Kuhn-Tucker conditions are necessary and sufficient:

$$L_{v_i} = p_i - \lambda f_{vi} = 0 \quad (i = 1, \dots, n) \quad [C.7]$$

$$L_{k_j} = -\lambda f_{kj} + \mu_j = 0 \quad (j = 1, \dots, m) \quad [C.8]$$

$$L_\lambda = y - f(z_v, z_k) = 0 \quad [C.9]$$

$$L_{\mu_j} = z_{kj} - z_{kj}^0 \leq 0, \quad \mu_j \geq 0, \quad \mu_j (z_{kj} - z_{kj}^0) = 0 \quad (j = 1, \dots, m) \quad [C.10]$$

The conditions [C.7] on the variable inputs are identical in form to those from the long-run problem [B.1] and have the same interpretation. The marginal rate of technical substitution between variable inputs will equal the ratio of their prices. The Lagrangean multiplier λ on the output constraint again gives the rate at which the objective function increases with y , only now λ is the short-run marginal cost rather than the long-run marginal cost.

μ_j is the Lagrange multiplier on the constraint on the amount of fixed input j and is the rate at which the objective function falls as the constraint is relaxed. (Note that z_{kj}^0 enters negatively in L whereas y enters positively.) It is the reduction in cost of producing y if the firm was given a free unit of the j th fixed input. From [C.8] we see that $\mu_j > 0$ only if the marginal product of the j th fixed input is positive at the solution. If the marginal product is zero then cost cannot be reduced by substituting the fixed input for variable inputs because output would fall below the required level. Using the Envelope Theorem (section 2J) the effect on the firm's cost of being able to buy another unit of the j th fixed input at price p_{kj} is $p_{kj} - \mu_j$. Thus, if $p_{kj} < \mu_j$, the firm can reduce its cost by buying the additional unit of z_{kj} and reducing the amount of z_{vi} used.

The cost-minimizing variable and fixed-input vectors are $z_v(p, y, z_k^0)$ and $z_k(p, y, z_k^0)$, where $p = (p_v, p_k)$ is the $n + m$ vector of all input prices, and the short-run or restricted cost function is

$$S(p, y, z_k^0) = p_v z_v(p, y, z_k^0) + p_k z_k(p, y, z_k^0) \quad [C.11]$$

It possesses the same properties as the long-run cost function $C(p, y)$, as the reader should check (see Question 5). In particular Shephard's lemma holds for the variable inputs:

$$\frac{\partial S(p, y, z_k^0)}{\partial p_{vi}} = z_{vi}(p, y, z_k^0) \quad [C.12]$$

We can use Shephard's lemma to examine the relationship between the long and short run responses of input use to changes in input price. Let $z(p, y) = (z_v(p, y), z_k(p, y))$ be the $n + m$ input vector which solves the long-run cost-minimization problem at prices p for output y . Suppose that the fixed input vector z_k^0 in the short-run problem would be optimal in the long-run cost minimization problem for output y^0 at some input price vector p^0 so that $z_k(p^0, y^0) = z_k^0$. Then at prices p^0 the solutions of the short- and long-run problems of minimizing the cost of producing y^0 are identical. To see this, note that $(z_v(p^0, y^0), z_k(p^0, y^0))$ solves the long-run problem if and only if, for all feasible (z_v, z_k^0) in

$Z(y^0)$

$$p_v^0 z_v(p^0, y^0) + p_k^0 z_k(p^0, y^0) \leq p_v^0 z_v + p_k^0 z_k^0$$

But by assumption $z_k(p^0, y^0) = z_k^0$ and so [C.13] implies

$$p_v^0 z_v(p^0, y^0) + p_k^0 z_k^0 \leq p_v^0 z_v + p_k^0 z_k^0$$

so that $(z_v(p^0, y^0), z_k^0)$ also solves the short-run problem. Hence we have

$$z_v(p^0, y^0, z_k^0) = z_v(p^0, y^0)$$

and

$$S(p^0, y^0, z_k^0) = C(p^0, y^0)$$

At other prices p the long- and short-run cost-minimizing z_v will not coincide and the definition of $(z_v(p, y^0), z_k(p, y^0))$ as the long-run cost-minimizing choice for output y^0 implies that

$$C(p, y^0) = p_v z_v(p, y^0) + p_k z_k(p, y^0)$$

$$\leq p_v z_v(p, y^0, z_k^0) + p_k z_k^0 = S(p, y^0, z_k^0)$$

[C.17]

In part (a) of Fig. 8.12 the short- and long-run cost functions are plotted against one of the variable input prices, all other prices being held constant at p_{vj}^0 ($j = 1, \dots, n; j \neq i$)

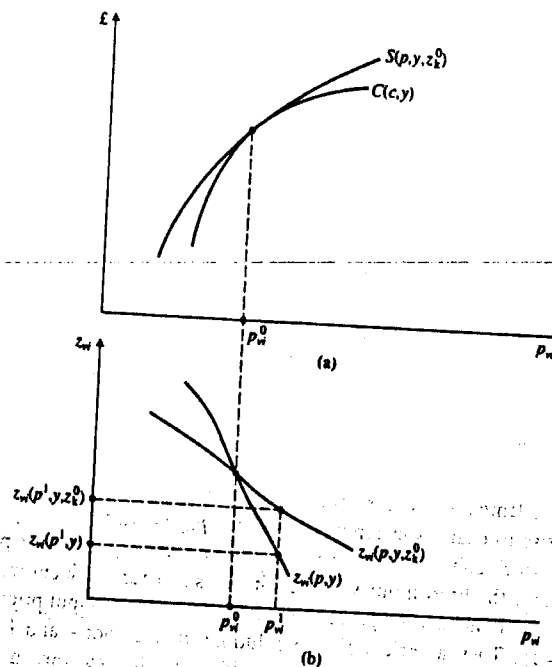


Fig. 8.12

[C.13]

[C.14]

[C.15]

[C.16]

or p_{vj} ($j = 1, \dots, m$) and output being held constant at y^0 . $S(p, y^0, z_k^0)$ lies above $C(p, y^0)$ everywhere except at $p_{vi} = p_{vi}^0$ where p is then equal to p^0 . In the neighbourhood of p_{vi}^0 , $S(p, y^0, z_k^0)$ must be flatter than $C(p, y^0)$ for $p_{vi} < p_{vi}^0$, steeper than it for $p_{vi} > p_{vi}^0$ and tangent to it at $p_{vi} = p_{vi}^0$. But the slopes of S and C in part (a) are just their derivatives with respect to p_{vi} and Shephard's lemma holds for the variable inputs in both the short- and the long-run. Hence $S_{vi}(p, y^0, z_k^0) = z_{vi}(p, y^0, z_k^0)$ is smaller than, greater than or equal to $C_{vi}(p, y^0) = z_{vi}(p, y^0)$ as p_{vi} is less than, greater than or equal to p_{vi}^0 . The vertical axis of part (b) of Fig. 8.12 plots the slopes of S and C with respect to p_{vi} against p_{vi} . Comparing, for example, the effects on $z_{vi}(p, y^0, z_k^0)$ and $z_{vi}(p, y^0)$ of an increase in p_{vi} from p_{vi}^0 to p_{vi}^1 , we see that in the neighbourhood of p^0 the response of the cost-minimizing demand for z_{vi} to changes in its price is smaller in the short- than in the long-run problem:

$$\frac{\partial z_{vi}(p^0, y^0, z_k^0)}{\partial p_{vi}} < \frac{\partial z_{vi}(p^0, y^0)}{\partial p_{vi}} \quad [C.18]$$

This result illustrates the *Le Chatelier-Samuelson Principle* that imposing additional constraints on an optimization problem will reduce the responsiveness of choice variables to changes in exogenous variables. The two-input case is an extreme example: in the short run the cost-minimizing input mix is not affected by the input prices, whereas in the long-run problem we would expect that choices do vary with p .

A similar envelope argument can be used to confirm our earlier diagrammatic analysis of the relationship between short- and long-run curves. Instead of comparing the effect on short- and long-run cost of varying input prices while holding output constant we would compare the effects of varying output while holding prices constant. At p^0 the short- and long-run costs of producing y^0 are equal: $C(p^0, y^0) = S(p^0, y^0)$ but at other outputs

$$C(p^0, y) = p_v^0 z_v(p^0, y) + p_k^0 z_k(p^0, y) \leq p_v^0 z_v(p^0, y, z_k^0) + p_k^0 z_k^0 = S(p^0, y, z_k^0)$$

The short-run cost function lies above the long-run function at all outputs except at y^0 where it is tangent to it. Thus short-run marginal cost will equal long-run marginal cost at y^0 and increase more rapidly with output than long-run marginal cost in the neighbourhood of y^0 .

Exercise 8C

1. Solve the short-run cost minimization problem and draw the short-run cost curves for a multiprocess fixed proportions technology. Why does the short-run marginal cost curve become vertical?
2. Repeat Question 1 for the case of a Cobb-Douglas production function. Does the SMC curve become vertical? Why, or why not?
3. What happens in Fig. 8.8 if part of the ridge line lies below the horizontal line at z_2^0 ? How will the short-run expansion path and cost curves differ?
4. Assume that the firm wishes to produce a given output next month, has already contracted to hire z_2^0 units of labour at a price of p_2 per unit and cannot fire

workers without giving them a month's notice, i.e. without paying them for the time they would have worked during the month. Additional labour can, however, be hired for next month at a price of p_2 , though the firm cannot resell the labour hours it has already contracted for. Solve the short-run cost minimization problem for a firm with one other freely variable input and draw the short-run cost curves. How do the results obtained differ from those in the text?

- 5.* Show that the short-run cost function $S(p, y, z_k^0)$ derived from [C.5] satisfies properties (a) to (d) in section B.

D. Cost minimization with several plants

Many firms possess more than one plant capable of producing their product and hence face the problem of allocating a required total output among their plants so as to minimize the cost of producing that output. The problem can be solved in two stages. First, each plant solves the problem of producing a given output level at least cost in that plant, subject to the production function for that plant, by choosing a plant cost-minimizing input bundle. Each plant then has a cost function derived in the usual way. In the two-plant problem the plant cost function is

$$C_i = C_i(y^i) \quad (i = 1, 2)$$

where C_i is total cost in plant i , y^i is the output in plant i (y^1 and y^2 are the same goods but produced in different plants) and the input prices have been omitted from the cost functions. C_i may be the short- or long-run cost function depending on the constraints on the adjustment of inputs. The second stage of the problem is

$$\min_{y^1, y^2} C = C_1(y^1) + C_2(y^2) \quad \text{s.t.} \quad (i) \quad y^1 + y^2 \geq y^0 \quad (D.1)$$

$$(ii) \quad y^i \geq 0 \quad (i = 1, 2)$$

The marginal cost in plant i is $C'_i(y^i)$ and we assume that the cost functions are strictly convex in y^i so that marginal cost is increasing with output: $C''_i(y^i) > 0$, $y^i \geq 0$. This means that $C_1 + C_2$ is convex in the output levels and thus the Kuhn-Tucker conditions are sufficient as well as necessary. The Lagrangean is

$$L = C_1(y^1) + C_2(y^2) + \lambda(y^0 - y^1 - y^2) \quad (D.2)$$

and the Kuhn-Tucker conditions are

$$L_i = C'_i(y^i) - \lambda \geq 0, \quad y^i \geq 0, \quad y^i[C'_i(y^i) - \lambda] = 0, \quad (i = 1, 2) \quad (D.3)$$

$$L_\lambda = y^0 - y^1 - y^2 \leq 0, \quad \lambda \geq 0, \quad \lambda(y^0 - y^1 - y^2) = 0 \quad (D.4)$$

λ is the rate at which the firm's cost would increase if its output requirement y^0 was increased: it is the marginal cost of the multi-plant firm. At least one of the y^i must be positive to satisfy the output requirement constraint and for the positive y^i it must also be true that $\lambda = C'_i(y^i)$. Since marginal cost is positive so must be λ and so the output requirement constraint must bind at the solution. Unsurprisingly, a cost minimizing firm with positive marginal cost will never produce more output than it requires.

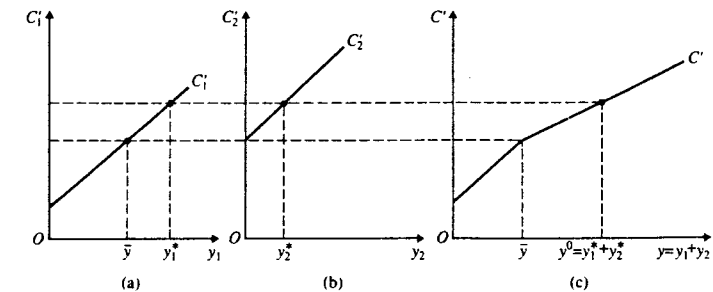


Fig. 8.13

There are two types of solution depending on whether only one or both plants are used when costs are minimized. If both plants are used then [D.3] implies

$$C'_1(y^1) = C'_2(y^2) = \lambda$$

and costs are minimized when output is allocated between the plants to equalize marginal costs in the two plants. Figure 8.13 illustrates this type of solution. The marginal cost curves C'_1 and C'_2 for the two plants are shown in parts (a) and (b) and the cost minimizing output of plant i is y^{i*} , with $y^{1*} + y^{2*} = y^0$. If C'_1 was not equal to C'_2 at an allocation where both plants are used it would be possible to reduce cost by transferring output from the plant with the higher marginal cost to the plant with a lower marginal cost. For example if $C'_1 > C'_2$ increasing y^2 by one unit and reducing y^1 by one unit would leave total output unchanged and reduce total cost by approximately $C'_1 - C'_2 > 0$.

The other type of solution has only one of the plants in operation. Suppose that $C'_2(0) > C'_1(y^0)$. Then it is cost minimizing to use only plant 1. Marginal costs are not equalized by transferring output from the high marginal cost plant 2 to the low marginal cost plant 1 because plant 2 output cannot be reduced below zero. In terms of [D.3] we have $C'_2(0) - \lambda = C'_2(0) - C'_1(y^0) > 0$ which implies from the complementary slackness condition that the optimal level of $y^{2*} = 0$. In Fig. 8.13 the firm would produce a required output of less than \bar{y} (defined by $C'_2(0) = C'_1(\bar{y})$) only in plant 1, leaving plant 2 idle.

Part (c) of the figure shows the firm's marginal cost of producing different total outputs, given that at each output it allocates output between the two plants so as to minimize total cost. For outputs of \bar{y} or less only plant 1 is used and so the firm's marginal cost curve $C'(y)$ is just the marginal cost curve of plant 1. For outputs of more than \bar{y} cost minimization requires that both plants are used and that plant marginal costs are equal. The firm's marginal cost curve is then the horizontal sum of the marginal cost curves of the two plants.

Least cost production with increasing returns: 'natural' monopoly

If plants all have identical strictly convex cost functions, least cost production requires that each produces the same amount (so that marginal costs are equalized) whatever the total output required. When the cost functions are not convex this conclusion may not be valid and it may be most efficient, in the sense of producing a given output at least cost,

to produce all the output in one of the identical plants. The fact that it is cheaper to produce output in one plant rather than in several obviously has implications for the number of firms in the market. When it is cost minimizing to produce any output up to y^0 in one plant there is said to be 'natural' monopoly in that output. The implication is that with this type of technology one would expect to see only one firm producing the entire industry output. However, a satisfactory theory of the equilibrium number of firms in an industry must rest on more than the technology: the entry and output decisions of profit maximizing firms depend on the revenues they anticipate from different decisions as well as their costs. Hence the quotation marks. The reader should remember in what follows that the relationship between the properties of the cost function and the cost minimizing number of firms or plants is not a complete explanation of monopoly, although it may be an important part of such an explanation.

The cost function for the identical plants is denoted $C(y)$ and there is 'natural' monopoly for $y \leq y^0$ when

$$C(y^1 + y^2) < C(y^1) + C(y^2) \quad (0 \leq y^1 + y^2 \leq y^0) \quad [\text{D.5}]$$

$C(y^1 + y^2)$ is the cost of producing $y^1 + y^2$ in a single plant, $C(y^i)$ the cost of producing y^i in a single plant. If [D.5] holds it is cheaper to produce a total output of $y^1 + y^2$ in a single plant rather than using two identical plants to produce separately outputs of y^1 and y^2 . A cost function which satisfies [D.5] is *sub-additive* so that sub-additivity and 'natural' monopoly are merely different labels for the same type of cost function. We prefer sub-additivity since it is a more neutral term.

We can establish some relationships between sub-additivity and other properties of the cost function:

(a) We have already seen that if $C(y)$ is strictly convex ($C''(y) > 0$ for all $y \geq 0$) [D.5] cannot hold (just apply the discussion of [D.3] with $C_1(y) = C_2(y)$). *Sub-additivity requires some degree of non-convexity in the cost function.*

(b) If there are economies of scale the average cost of production falls with output (see section B) and so

$$\frac{C(y^1 + y^2)}{y^1 + y^2} < \frac{C(y^i)}{y^i} \quad (i = 1, 2) \quad [\text{D.6}]$$

must hold if $y^1 > 0, y^2 > 0$ (so that $y^i < y^1 + y^2$). Multiplying both sides of the inequalities by y^i and adding the two inequalities we have

$$\frac{y^1 C(y^1 + y^2)}{y^1 + y^2} + \frac{y^2 C(y^1 + y^2)}{y^1 + y^2} = C(y^1 + y^2) < C(y^1) + C(y^2) \quad [\text{D.7}]$$

and so we have established that *economies of scale imply sub-additivity*.

(c) Unfortunately the converse does not hold and *sub-additivity over an output range does not imply economies of scale over that output range*. Figure 8.14 gives an example of a cost function with a U-shaped average cost curve. There is a fixed cost F which must be incurred to produce any output but which can be avoided if no output is produced. The cost function is discontinuous at $y = 0$ and so it is not everywhere convex despite the fact that marginal cost (the slope of $C(y)$) is increasing at all positive output levels. Average cost $C(y)/y$ is measured by the slope of a line from the origin to the cost function and

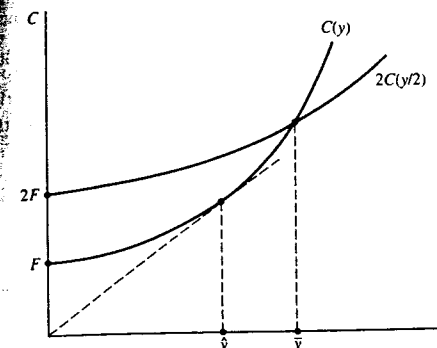


Fig. 8.14

up to output \hat{y} average cost decreases with output (economies of scale) and beyond \hat{y} average cost increases with y (diseconomies of scale). For positive outputs marginal cost increases with output so that if production is carried on in two plants it is cost minimizing to produce a total output of y by producing $y/2$ in each plant. Hence with two-plant production total cost is $2C(y/2)$. It is clear from the diagram that for output less than \hat{y} it is cheaper to produce in one plant only because of the saving on fixed costs. (If there were no fixed costs (so that $C(y)$ is shifted down by F to start at the origin) two plants would be more efficient than one because of the saving on variable costs achieved by equalizing marginal costs.) Note that over the range $\hat{y} \leq y \leq \bar{y}$ there are diseconomies of scale but it is cheaper to use one plant rather than two so that sub-additivity can occur even with diseconomies of scale.

Exercise 8D

- Merit order.** Suppose that a power company has n different plants each of which embodies a different fixed proportions process. Each plant has a maximum output rate which cannot be exceeded in the period because of constraints on the fixed input in each plant. Derive the short-run marginal cost curve for the firm and the merit order of plants which shows the order in which the plants are brought into production as the required output level increases.
- Cost minimization with U-shaped cost curves.** Suppose that the firm has two plants with identical cost functions $C(y) = F + V(y) = F + vy^\alpha$, ($F > 0, v > 0$), $C(0) = 0$.
 - Why may the cost minimization conditions [D.3], [D.4] fail to identify the least cost allocation?
 - Over what output ranges are there economies of scale when $\alpha = 1$ and when $\alpha = 3$?
 - Over what output ranges is it cheaper to use one rather than two plants (sub-additivity) when $\alpha = 1$ and when $\alpha = 3$?

E. Multi-product cost functions*

If the firm produces two outputs, y_1 and y_2 , its problem is to minimize the cost of producing specified levels, y_1^0 and y_2^0 of its products. The production function constraint is written in the implicit form $g(y_1, y_2, z_1, \dots, z_n) \leq 0$ of section 7D. If input prices are positive the firm will produce exactly the specified levels of outputs ($y_1 = y_1^0, y_2 = y_2^0$) and in a technically efficient way: $g(\dots) = 0$. The cost minimization problem therefore is

$$\begin{aligned} \min_{z_1, \dots, z_n} \sum p_i z_i \text{ s.t. } & \text{(i) } g(y_1^0, y_2^0, z_1, \dots, z_n) = 0 \\ & \text{(ii) } z_i \geq 0 \quad (i = 1, \dots, n) \end{aligned} \quad [\text{E.1}]$$

The Lagrange function is

$$L = \sum p_i z_i + \lambda g(y_1^0, y_2^0, z_1, \dots, z_n)$$

and the first-order conditions on the inputs are, in an interior solution

$$\frac{\partial L}{\partial z_i} = p_i + \lambda g_i = 0 \quad (i = 1, \dots, n) \quad [\text{E.2}]$$

Writing the conditions as $p_i = -\lambda g_i$ and dividing the i th condition by the j th gives

$$\frac{p_i}{p_j} = \frac{g_i}{g_j} \quad [\text{E.3}]$$

In section 7D it was demonstrated that g_i/g_j is the marginal rate of technical substitution between the two inputs so that the necessary condition for cost minimization in the multi-product case is identical with that in the single-product case.

The Lagrange multiplier λ has a somewhat different interpretation in the multi-product problem [E.1]. λ is attached to the production function constraint rather than to the output constraint as in the single output problem. It measures the rate at which the minimized cost of production is reduced if the production function constraint is relaxed slightly, i.e. if it is possible to produce the specified outputs with smaller inputs.

The cost function and joint costs

As in the single output case the optimal input levels will be functions of input prices and the required output levels:

$$z_i^* = z_i(y_1, y_2; p)$$

and substitution in $\sum p_i z_i$ gives the multi-product cost function which shows the minimized cost of production as a function of the output levels and input prices:

$$C = \sum p_i z_i^* = C(y_1, y_2; p) \quad [\text{E.4}]$$

The multi-product cost function possesses all the convenient properties (a)–(d) of the single product cost function in section B. The arguments are very similar so we leave them to the reader.

Joint production and the cost function

Part of the explanation for the fact that multi-product firms are more common than single-product firms is that in some circumstances production of several different goods in the same plant or firm is less costly than if the same quantities of the different goods were produced in specialist single product firms. The relationship between cost and output for multi-product cost functions is therefore of some interest in explaining the existence of multi-product firms. The marginal cost of good i is just the partial derivative of [E.4] with respect to good i : $C_i(y_1, y_2, p) = \partial C(y_1, y_2, p) / \partial y_i$. In section B there were said to be economies of scale if the elasticity of cost with respect to output was less than one. In the multi-product case we can examine the effect on cost of an equal proportionate increase in all its outputs. Thus there are *multi-product economies of scale* if the elasticity of cost with respect to the scale of output, E_t^C , defined as

$$E_t^C = \frac{\partial C(ty_1, ty_2, p)}{\partial t} \cdot \frac{t}{C(ty_1, ty_2, p)} = \sum_i C_i y_i \cdot \frac{t}{C} \quad [\text{E.5}]$$

is less than one. For given y increases in the output scale parameter t imply equal proportional increases in all outputs and so t can be thought of as a measure of size of the firm's output. The last term in [E.5] is the reciprocal of $C(ty_1, ty_2, p)/t$ which can be interpreted as a kind of average cost since it divides cost by a measure of output. It is known as the *ray average cost* since increases in t correspond to movements along a ray from the origin in output space. The term $\sum_i C_i y_i$ in [E.5] is the rate of change of cost as the firm increases its output scale and can be defined as *ray marginal cost*. [E.5] is therefore rather similar to the elasticity of cost with respect to output in the case of a single product firm.

To see when joint production is less costly than specialist production at some input price vector p define the *stand-alone cost* of y_1 as the minimized cost of producing y_1 when $y_2 = 0$.

$$C(y_1, 0, p) = C^1(y_1, p) \quad [\text{E.6}]$$

and analogously for the stand-alone cost of good 2. Joint production is less costly than specialist production if

$$C(y_1, y_2, p) < C(y_1, 0, p) + C(0, y_2, p) \quad [\text{E.7}]$$

Conversely, if the inequality in [E.7] is replaced with an equality the cost function is *output separable* and any output vector $y = (y_1, y_2)$ could be produced as cheaply in separate specialist firms as in a multi-product firm.

If [E.7] holds for all output vectors $0 \leq y \leq y^0$ the cost function is said to exhibit *economies of scope* over this range. Since input prices are assumed to be independent of the firm's decisions, whether [E.7] holds or not depends on the form of its production function. The cost of producing, say, good 1 is unaffected by the output of good 2 only if the inputs required to produce good 1 do not vary with the output of good 2, that is if the production function is separable. When y_1 and y_2 are joint products in the sense of section 7D the cost function is non-separable. Obvious examples in which it is cheaper to produce a pair of goods in one organization rather than in two range from beef and

cow-hides to peak and off-peak electricity. (What about research and teaching in a university?)

In section D we introduced the concept of sub-additivity ('natural' monopoly) in the context of a single type of good but the same issues arise when plants or firms can produce more than one type of good. We can extend the definition of sub-additivity to the multi-product case by saying that the cost function $C(y, p)$ is sub-additive if

$$C(y^1 + y^2, p) < C(y^1, p) + C(y^2, p) \quad [\text{E.8}]$$

where $y^i = (y_1^i, y_2^i)$ is an output vector. Our definition of sub-additivity in section D is just a special case of this with, say $y^i = (y_1^i, 0)$. When [E.8] holds for all $0 \leq y \leq y^0$ the cost function is *globally sub-additive* over this range and it is cheaper to organize production in a single firm or plant than in separate specialist production units.

The relationships between the economies of scope and scale and sub-additivity can turn out to be rather surprising. Intuition would suggest that if the cost function has multi-product economies of scale or economies of scope it would be cheaper to organize production in one unit rather than in separate production units. Unfortunately, this is not so: a cost function with economies of scope and multi-product economies of scale need not be sub-additive. Consider the following cost function (due to Sharkey, 1982):

$$C(y_1, y_2, p) = \begin{cases} a(p)y_2 + b(p)y_1^2/y_2 & \text{for } y_1 \leq ky_2 \\ a(p)y_1 + b(p)y_2^2/y_1 & \text{for } y_1 \geq ky_2 \end{cases} \quad [\text{E.9}]$$

where $a(p)$, $b(p)$, k are positive coefficients. The reader should check that this function has both multi-product economies of scale and economies of scope. However, if $a = b = k = 1$ we have $C(3, 3, p) = 6$, $C(1, 2, p) = 2.5$, and $C(2, 1) = 2.5$ so that C is not globally sub-additive. The reason economies of scale and scope do not imply global sub-additivity is that output vectors used to define economies of scale and scope are highly restricted. With economies of scale one is examining the effects on cost of movements along rays from the origin in output space and with economies of scope one is comparing the cost of a vector (y_1, y_2) with the costs of $(y_1, 0)$ and $(0, y_2)$, i.e. with the costs of projections of the vector on to the y_1 and y_2 axes. Global sub-additivity, however, requires comparisons of the cost of (y_1, y_2) with the costs of all vectors which add up to (y_1, y_2) , not just those on the ray from y_1, y_2 to the origin or those at $(y_1, 0)$ and $(0, y_2)$.

One condition on the cost function which implies sub-additivity is *cost complementarity*:

$$C(y^1 + y^2 + y^3, p) - C(y^1 + y^2, p) < C(y^1 + y^3, p) - C(y^1, p) \quad [\text{E.10}]$$

for all output vectors $y^1 \geq 0$, $y^2 > 0$, $y^3 > 0$. If [E.10] holds the incremental cost arising from increasing output by the vector y^3 is smaller the larger is the initial output vector. By considering special cases in which $y^3 = (\Delta y_1, 0)$ or $y^3 = (0, \Delta y_2)$ and $y^2 = (\Delta y_1, 0)$ or $y^2 = (0, \Delta y_2)$ and taking appropriate limits we can show that [E.10] is equivalent to increases in good j reducing the marginal cost of good i : $C_{ij} = \partial^2 C(y_1, y_2, p) / \partial y_i \partial y_j < 0$ ($i, j = 1, 2$). It can be shown that *cost complementarity implies economies of scope and multi-product economies of scale* (see the exercises). More importantly *cost complementarity implies global sub-additivity*. To see this just use the definition [E.10] with $y^1 = (0, 0)$, so that $C(y^1, p) = C(0, 0, p) = 0$ and [E.10] becomes

$$C(y^2 + y^3, p) - C(y^2, p) < C(y^3, p)$$

which rearranges to give [E.8].

Sub-additivity has implications for the way in which cost minimizing firms will organize production. It also suggests that attempts to allocate the total cost of producing several products among the products so as to yield a 'cost' of producing each particular product will be meaningless. Accounting conventions may apportion the total cost $C(y_1, y_2, p)$ between the two products by various procedures but the resulting relationship between outputs and cost provides no information useful for decision-making. Any attempt, for example, to decentralize production by creating product divisions and instructing them to maximize the difference between their revenue and the 'cost' allocated to their product will lead to sub-optimization. Similarly, attempts to regulate the behaviour of public utilities on the basis of costs apportioned between different products may lead regulators astray. Sensible decisions require information about the effects of a change in say output 1 on the total costs of the firm, not on the 'costs' apportioned to that product by arbitrary conventions. We return to this question in the next chapter but note that apportioning cost between different products will only be sensible if the cost function is output separable.

Exercise 8E

1. *Concavity and sub-additivity.* Consider the cost function $C(y_1, y_2, p) = a_1(p)\sqrt{y_1} + a_2(p)\sqrt{y_2} + b(p)\sqrt{(y_1 y_2)}$ with $a_1(p)$, $a_2(p)$, $b(p)$ equal to 1. Does it exhibit (a) concavity in output; (b) multi-product economies of scale; (c) economies of scope; (d) cost complementarity; (e) global sub-additivity?
- 2.* Show that cost complementarity implies (a) economies of scale and (b) economies of scope.

References and further reading

The relationship between cost and production functions is surveyed in

R. G. Chambers. *Applied Production Analysis: A Dual Approach*, Cambridge University Press, Cambridge, 1988,

and at a very rigorous level in

R. W. Shephard. *Theory of Cost and Production Functions*, Princeton University Press, Princeton, NJ, 1970.

There are many illustrations of the use of the theory of multi-product cost functions in the analysis of industrial structure and the regulation of firms in

W. J. Baumol, J. C. Panzar and R. D. Willig. *Contestable Markets and the Theory of Industrial Structure*, Harcourt Brace Jovanovich, New York, 1982.

W. W. Sharkey. *The Theory of Natural Monopoly*, Cambridge University Press, Cambridge, 1982.

D. F. Spulber. *Regulation and Markets*, MIT Press, Cambridge, MA, 1989.

Adjustment costs are considered in

S. J. Nickell. *The Investment Decision of Firms*, Cambridge University Press, Cambridge, 1978.

Rather different concepts of cost to that developed in this chapter are examined in

A. A. Alchian. 'Costs and outputs', in M. Abramovitz *et al.* (eds), *The Allocation of Economic Resources*, Stanford University Press, Stanford, CA, 1959.

J. M. Buchanan. *Cost and Choice: An Enquiry into Economic Theory*, Markham, Chicago, 1969.

CHAPTER 9

Supply

The discussion of the technological constraints on the firm in Chapter 7 required no mention of the firm's objectives and even for the derivation of the cost functions and curves of Chapter 8 all that was required was the assumption that the firm wished to produce each output level at least cost. Nothing was said about how that output level was determined. It is now necessary to make some assumptions about the objectives of the firm. We can then proceed to analyse the firm's output decision and its responses to changes in the environment. The assumption we adopt is that the firm wishes to maximize its profits. This assumption has not gone unchallenged, as we saw in Chapter 6, and the implications of some suggested alternatives are considered in Chapter 13.

The existence of adjustment costs means that the firm must make two kinds of decisions at any point in time: it must choose an output level that it will produce in the current period and it must *plan* the outputs to be produced in future periods. This plan of future outputs will imply a sequence of future input levels and this in turn will imply a programme of actions by the firm, to be implemented over time, beginning in the current period, to increase or decrease input levels to the planned future levels.

As in Chapter 8, we will not analyse this problem in its full generality but will instead consider a two-period approximation to it. At the start of period 0 the firm will choose (a) an output level for the current period (period 0) given the constraints on the adjustment of the fixed input and (b) a planned output level for period 1, given that all inputs are variable. Problem (a) is the short-run profit maximization problem which is analysed in section B and problem (b) the long-run profit maximization problem analysed in section A.

Sections A and B are concerned with the case of a single output y and two inputs z_1 and z_2 . Section C extends the analysis to a multiproduct firm. The firm is assumed to operate in competitive markets in the sense that it takes the prices of inputs and outputs as unaffected by its decisions. We discuss the implications of the firm's decisions affecting the prices it faces in Chapter 11 (monopoly) and section 15C (monopsony). When, as in this chapter, the firm treats prices as parameters, the maximum profit it can earn is a function of the prices it faces. In section D we discuss the properties of this maximum profit function and show that it is a useful tool for investigating the firm's behaviour.

A. Long-run profit maximization

The firm's long-run decision problem is to *plan* an output and input combination to maximize profit, π , where profit is revenue $R = py$ minus cost $\sum p_i z_i$, and p, p_i are the prices of y and z_i respectively. Formally, the problem is:

$$\begin{aligned} \max_{y, z_1, z_2} \pi &= py - \sum p_i z_i \quad \text{s.t. } y \leq f(z_1, z_2) \\ y &\geq 0, \quad z_1 \geq 0, \quad z_2 \geq 0 \end{aligned} \quad [\text{A.1}]$$

This problem can be reformulated in two equivalent ways:

(a) For any output, profit cannot be maximized unless cost is minimized. Hence, we can make use of the earlier analysis of cost minimization, and work with the long-run cost function $C(p_1, p_2, y)$, derived there. The profit maximization problem can be expressed as:

$$\max_y py - C(p_1, p_2, y) \quad \text{s.t. } y \geq 0 \quad [\text{A.2}]$$

The firm simply chooses the output level which maximizes profit given its revenue function py and cost function C .

This two-stage optimization procedure (minimizing costs to derive the cost function and then maximizing the difference between the revenue and cost functions) has thus reduced the profit maximization problem to a single decision variable problem and, as we will see below, this makes the analysis of the model based on this problem fairly easy.

(b) Alternatively, we can state the problem as follows: since prices are positive the profit maximizing firm will never produce in an output-inefficient way. If $y < f(z_1, z_2)$ then either y can be increased holding z_1 and z_2 constant, or one or both of the inputs can be reduced with y constant, and so profit cannot be at a maximum. Hence the production constraint on a profit maximizing firm can be written as $y = f(z_1, z_2)$. Since a choice of z_1 and z_2 determines y , there are only two independent decision variables: the two input levels. The firm's profit maximization problem is therefore:

$$\max_{z_1, z_2} \pi = p \cdot f(z_1, z_2) - p_1 z_1 - p_2 z_2 \quad \text{s.t. } z_1 \geq 0, \quad z_2 \geq 0 \quad [\text{A.3}]$$

We will use approach (b) in section 15A because it is useful when we wish to focus on the firm's input demands. In this section we use approach (a) to emphasize the firm's output decision.

Differentiating [A.2] with respect to y we have the following first order condition for y^* to provide a maximum of the profit function:

$$\frac{d\pi}{dy} = p - \frac{\partial C}{\partial y} \leq 0, \quad y^* \geq 0, \quad y^* \cdot \frac{d\pi}{dy} = 0 \quad [\text{A.4}]$$

Since nothing has been assumed about the shape of the profit function, [A.4] is a necessary but not a sufficient condition for y^* to yield a maximum. [A.4] may be satisfied by a number of local maxima or minima as Fig. 9.1 illustrates. The total cost, revenue and profit functions are plotted in part (a) and the marginal cost, revenue and profit and average cost functions in part (b).

It is clear from Fig. 9.1(a) that y^* is the global profit maximizing output and that y^* satisfies [A.4]. But consider two other output levels: $y = y^1$ and $y = 0$. At y^1 , $d\pi/dy = p - \partial C/\partial y < 0$ so that [A.4] is satisfied and this is a local profit maximum since profit is larger (loss is smaller) than at neighbouring feasible outputs. At y^1 , $d\pi/dy = 0$ but π is at a minimum. To distinguish between interior local maxima and minima (where $y > 0$) a second-order condition is required:

$$\frac{d^2\pi}{dy^2} = \frac{-\partial^2 C}{\partial y^2} < 0 \quad \text{i.e.} \quad \frac{\partial LMC}{\partial y} > 0 \quad [\text{A.5}]$$

This condition is satisfied at y^* but not at y^1 , and hence distinguishes between interior points ($y > 0$) which satisfy the necessary condition in [A.4] but which may be minima or maxima. Condition [A.5] is, however, *not applicable* at $y = 0$. The zero output position is a true local maximum because small permissible changes (i.e. increases) in y from $y = 0$ reduce profit (refer to Fig. 9.1(a)) even though LMC is falling at that point. We have in fact a case where there are multiple local optima and the global optimum can only be

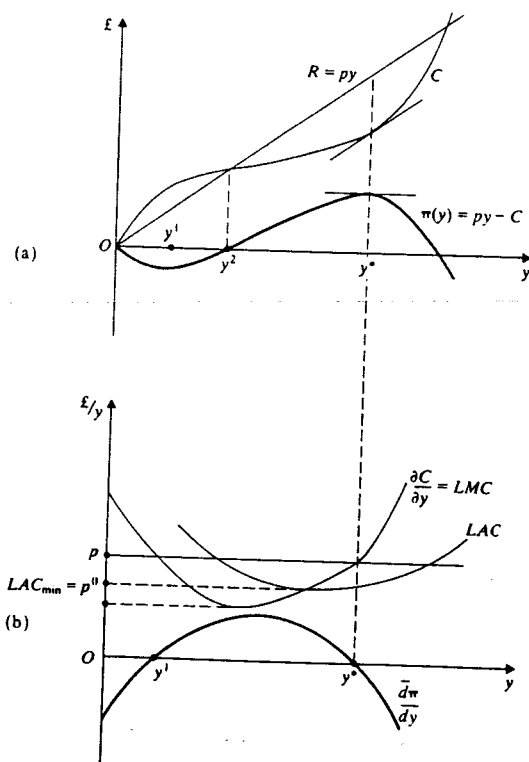


Fig. 9.1

and by direct comparison of these – profit or loss at $y = 0$ must be compared with profit loss at $y = y^*$. In the figure y^* is clearly superior, but it is easy to re-draw the curves in such a way that total cost is everywhere above total revenue and the interior point at which profit is maximized (loss is minimized) is inferior to $y = 0$ (draw the diagram).

In terms of the discussion of local and global optima in section 2D, the problem has arisen here because the conditions of the relevant theorem are not satisfied. The theorem states that if the feasible set is convex and the objective function is quasi-concave *every* local optimum is a global optimum, and so all local optima must yield equal values of the objective function. Here the feasible set defined by $y \geq 0$ is convex but the objective function is *not* quasi-concave. To see this, take two points at which profit is equal, say $y = 0$ and $y = y^2$ in Fig. 9.1(a) (where profit is zero). The definition of quasi-concavity requires that, for any pair of points at which profit is the same, the profit yielded by an output on the straight line joining them must be at least as great as that yielded by the two points. But, clearly, at all outputs on the straight line joining $y = 0$ and $y = y^2$ profit is less than zero and so the profit function is *not* quasi-concave. We cannot then be sure that every local maximum will be a global maximum and indeed we have just seen that, in the case shown in Fig. 9.1(a), one will not be.

Of course, as pointed out in section 2D, the conditions of the theorem are sufficient but not necessary. The reader is invited to re-draw Fig. 9.1(a) in such a way that $y = 0$ and $y = y^*$ are equally good. (Hint: look for a point of tangency.) The point is of course that this is a special case and in general we cannot guarantee that a local optimum is a global optimum when the profit function is not quasi-concave.

Long-run supply function

When $y^* > 0$ conditions [A.4] and [A.5] can be given a familiar interpretation. Condition [A.4] states that for profit to be maximized at y^* it is necessary that a small change in output adds as much to costs as it does to revenue. Marginal revenue (which is equal to the price of the product in a competitive market) must equal marginal cost. Condition [A.5] requires that marginal cost be increasing with output at y^* so that the marginal cost curve cuts the price line (the competitive firm's marginal revenue curve) from below. The firm maximizes profit by moving along its marginal cost curve until marginal cost is equal to price.

As Fig. 9.1 shows, the firm responds to an increase in the price of its output by moving along its LMC curve provided price exceeds long-run average cost. The portion of LMC curve above the LAC is therefore the *long-run supply curve* of the competitive firm.

More formally, the first-order condition

$$\pi_y(y^*; p, p_1, p_2) = p - C_y(p_1, p_2, y^*) = 0 \quad [\text{A.6}]$$

is an implicit function of y^* , p , p_1 and p_2 which can be solved to give the *long-run supply function of the competitive firm*:

$$y^* = y^*(p, p_1, p_2) \quad [\text{A.7}]$$

The fact that the firm increases y^* when p increases is clear from Fig. 9.1 but it is instructive to demonstrate this using the comparative static method of section 2I. A

change in p of dp must lead to a change dy^* in the firm's output if the firm is to continue to hold the change in output price must induce a change in output such that

$$d\pi_y(y^*; p, p_1, p_2) = \pi_{yy}(y^*; p, p_1, p_2) dy^* + \pi_{yp}(y^*; p, p_1, p_2) dp = 0 \quad [A.8]$$

where $\pi_{yp} = \partial\pi_y/\partial p$. Rearranging [A.8] gives

$$\frac{dy^*}{dp} = \frac{-\pi_{yp}}{\pi_{yy}} = \frac{1}{C_{yy}(p_1, p_2, y^*)} > 0 \quad [A.9]$$

(remember that the second-order condition [A.5] requires that $\pi_{yy} = -C_{yy} < 0$).

The firm's long-run supply decision will also depend on its cost conditions and Fig. 9.1 indicates any change in input prices or its technology which increases its long-run marginal cost will reduce output supplied at any given output price: the long-run supply curve of the firm will have shifted upward.

We can use the section 2I comparative static methodology to investigate the effects of an increase in the price of input i on the firm's supply decision. Totally differentiating [A.6] with respect to p_i and y^* and rearranging gives

$$\frac{dy^*}{dp_i} = \frac{-\pi_{yp_i}}{\pi_{yy}} = \frac{-C_{yp_i}(p_1, p_2, y^*)}{C_{yy}(p_1, p_2, y^*)} \quad [A.10]$$

where $C_{yp_i} = \partial C_y/\partial p_i$ is the effect of an increase in the price of input i on long-run marginal cost. Recalling from section 8B that C_{yp_i} is positive or negative as input i is normal or regressive we see that the firm's output is reduced or increased by an increase in price of input i as i is normal or regressive.

The firm's optimal y^* is zero if p is less than the minimum long-run average cost LAC_{\min} at which it can produce. The firm earns its maximum profit (of zero) by setting $y^* = 0$ if $p < LAC_{\min}$. In Fig. 9.1(b) an anticipated price of less than p^0 will cause the firm to plan to cease production next period since p^0 is the lowest price at which LAC can be covered.

The possibility that the optimal output can be zero means that our discussion of the firm's comparative static responses requires qualification. The firm's long run supply curve is the vertical axis (nothing is supplied) for $p < LAC_{\min}$ and its LMC curve for $p > LAC_{\min}$. If the firm has a U shaped LAC curve its supply curve will be discontinuous at $p^0 = LAC_{\min}$. At $p^0 = LAC_{\min}$ the firm would be indifferent between supplying $y^* = 0$ or the output at which LAC is minimized. Its long-run supply decision is then strictly speaking a correspondence rather than a function.

Exercise 9A

1. Show that the equilibrium conditions derived from problem [A.3] are equivalent to those from the two-stage approach to profit maximization.
2. Returns to scale and the supply function. Sketch the long-run supply function of a competitive firm with
 - (a) diminishing returns to scale
 - (b) constant returns to scale

- (c) why will the firm never plan to supply an output at which it has increasing returns to scale?

Input prices and the supply function. What is the effect of an increase in the price of input i on

- (a) LAC_{\min}
- (b) the output at which LAC is minimized?
- (c) sketch the effect of an increase in the price of input i on the firm's long run supply curve for (i) a normal input and (ii) a regressive input. (Hint: recall the analysis in section 8B.)

B. Short-run profit maximization

The firm's short-run problem is to choose output and input levels for the current period which maximize its current period profits, given that there are constraints on the adjustment of some of the inputs. Since inputs are chosen to minimize cost for any given output level the problem can be reduced to choosing current period output, y , to maximize the difference between revenue and short-run cost:

$$\max_y \pi = py - S(p_1, p_2, z_2^0, y) \quad \text{s.t. } y \geq 0 \quad [B.1]$$

where the constraint on the adjustment of z_2 is assumed to be an upper limit on the use of z_2 and the firm must pay for z_2^0 units irrespective of use. (See section 8C on the firm's short-run cost function.)

The first- and second-order conditions for this problem are very similar in form and interpretation to [A.4] and [A.5]. The firm will either produce where $p = \partial S/\partial y = SMC$ and where the SMC curve cuts the horizontal price line from below; or the firm will produce nothing if price is less than short-run average opportunity cost (average variable cost) at all positive outputs.

In the short-run the maximized level of profit may be negative, even if p exceeds minimum AVC . In Fig. 9.2, for example, which is based on Fig. 8.10 the firm makes a loss if

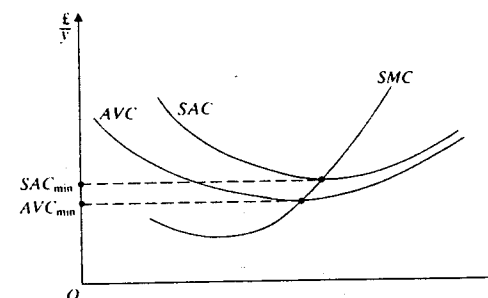


Fig. 9.2

$p < SAC_{\min}$ since fixed costs ($p_2 z_2^0$) are not covered. If p is less than AVC_{\min} the firm will set $y = 0$ since positive y implies that revenue would not cover variable cost, and so a loss would be made in addition to the loss on fixed costs. Conversely, if p exceeds AVC_{\min} then revenue is made over and above variable costs, so that some of the fixed costs are recovered by producing and selling some output. The firm may still make a loss but this is lower than the loss at zero output, which is equal to the fixed cost.

The firm's *short-run supply curve*, which shows the output it wishes to produce given the prevailing constraints on the adjustment of its inputs, will be the SMC curve for $p \geq AVC_{\min}$ and the vertical axis at $y = 0$ for $p < AVC_{\min}$. The firm's short-run supply curve, therefore, is discontinuous when the minimum of the AVC curve does not occur at $y = 0$.

The relationship between long- and short-run profit maximization

We pointed out in the introduction to this chapter, and in section 8A, that the firm makes two kinds of decisions at the start of each period: (a) it chooses the actual output level for that period, given the constraints on the adjustment of its inputs; (b) it *plans* an output level for the next period, when all inputs are freely variable (provided the decision to change them is made at the start of the current period). The first decision is the short-run, and the second the long-run, problem. We will now investigate in more detail how the two types of decision are related.

Some new notation is needed to distinguish between actual and planned, and between actual and forecast, magnitudes:

- y_a^t : actual output in period t .
- y_p^t : planned output in period t , decided upon in period $t - 1$.
- p_a^t : actual price of output in period t .
- p_f^t : forecast of price of output in period t made in period $t - 1$.

Since all inputs are freely variable after the current period, plans and expectations need only be made one period ahead, so that as indicated y_p^t refers to a plan made at period $t - 1$ and p_f^t to the firm's forecast of p_a^t made at period $t - 1$. It is assumed for simplicity that input prices and technological conditions are constant over all periods and that they are correctly anticipated at all times, so that actual and expected cost curves coincide and are the same in each period. To make the analysis more concrete let us take z_2 to be a measure of *plant size*.

Initially, at the start of period 0 the firm has a given plant size (z_2^0) which it cannot vary in period 0. Its short run cost curves are shown in Fig. 9.3 as SMC^0 and SAC^0 . In period 0 the firm maximizes its profits by equating short-run marginal cost to the known, current price of y (p_a^0) and so produces y_a^0 . At the same time the firm *plans* an output for period 1. Since the level of z_2 for period 1 can be varied if the decision to do so is made at the start of period 0, the relevant cost curves for planning the next period's output are the long-run curves LMC and LAC in Fig. 9.3. (Recall that these curves are derived from a cost minimization problem in which all inputs are freely variable.) At the start of period 0 the firm *expects* the period 1 price to be p_f^1 and so it *plans* to maximize period 1 profit by producing y_p^1 , where $p_f^1 = LMC$. The planned period 1 output in turn implies that the

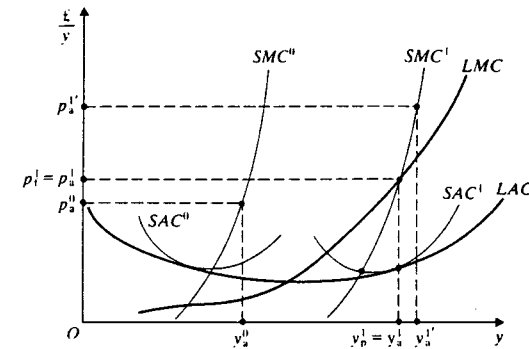


Fig. 9.3

period 1 level of z_2 is z_2^1 . To ensure that z_2^1 is actually available at the start of period 1, the firm must, at the start of period 0, order and install the additional plant required. Hence the decisions taken in period 0 are (a) to set the actual output in period 0, on the basis of the actual price p_a^0 and actual plant size z_2^0 ; (b) to choose the plant size z_2^1 for period 1, on the basis of planned period 1 output, which in turn depends on the forecast period 1 price.

At the start of period 1 the firm's actual plant is z_2^1 , giving rise to the short-run cost curves SMC^1 , SAC^1 . Suppose that the actual price is p_a^1 . Period 1 profit is maximized by equating SMC^1 to the actual price. In this case the firm's forecast was correct and $p_a^1 = p_f^1$. This means that actual and planned period 1 output are equal: $y_a^1 = y_p^1$. Note that at this output level $SMC^1 = LMC$, indicating that the actual plant (z_2^1) is the optimal plant for producing that output level.

The firm will also plan in period 1 for an output for period 2, based on its forecast p_f^2 of the period 2 price, and this will imply a period 1 decision on the actual plant for period 2 (z_2^2). If the firm expects p_f^2 to equal p_a^1 then $y_p^2 = y_a^1$ and there is no need to adjust plant size ($z_2^2 = z_2^1$). The firm will then be in *long-run equilibrium*: it will be maximizing profit for the current period (1) and its current plant will be optimal for the next period (2), given the firm's forecast of the next period's price (and ignoring depreciation).

Suppose, however, that at time 0 the firm had made the wrong price forecast, i.e. the actual and forecast period 1 prices differ (e.g. actual period 1 price is $p_a^1 > p_f^1$). The firm would find that its actual period 1 plant (z_2^1) was not optimal for the market price p_a^1 . In order to maximize period 1 profit, given z_2^1 and the corresponding SMC^1 curve, the firm will set $SMC^1 = p_a^1$ and produce the output y_a^1 . At the same time it will plan to produce y_p^1 , given its price forecast p_f^1 , and it will adjust its plant if $p_f^1 \neq p_a^1$, i.e. if its *forecast* of the price has changed (rather than if $p_a^1 \neq p_f^1$. Explain why.)

The output that the firm *plans* to produce in the next period, based on its forecast of the next period's price, determines the actual plant in the next period, but if the forecast is incorrect actual output next period will in general differ from that planned. The plan made commits the firm to a particular plant size next period, but *not* to a particular output level. When the firm chooses its current output at the start of a period it is *always* 'in the short-run': its plant size is fixed by the plan made in the previous period and is unalterable

in the current period. Hence the firm will *always* produce where $p'_a = SMC$ in order to maximize current period profit. If the past forecast was correct, then the current plant is optimal and the firm will be producing where $LMC = SMC^t = p'_a = p'_f$. If the past forecast was incorrect then the firm will not produce where $LMC = SMC^t$ and the existing plant will not be optimal. Short-run marginal cost and *actual* price determine *actual* output in the current period. Long-run marginal cost and the *forecast* price determine *planned* output and *actual* plant in the next period.

The relationship between forecast and actual, and planned and actual magnitudes can be represented in the following way:

$$\begin{array}{c} p'_f \rightarrow y'_p{}^{t+1} \rightarrow z'_2{}^{t+1} \rightarrow SMC^{t+1} \\ \quad \quad \quad \searrow \\ \quad \quad \quad y'_a{}^{t+1} \\ \quad \quad \quad \nearrow \\ p'_a{}^{t+1} \rightarrow p'_f{}^{t+2} \rightarrow y'_p{}^{t+2} \rightarrow z'_2{}^{t+2} \rightarrow \dots \end{array}$$

This emphasizes that a model which attempts to predict the firm's *actual* behaviour must include a sub-model of the way in which the firm makes its price forecasts. In the diagram above, for example, the dashed line from actual price to the forecast of the next period's price indicates that the forecast may depend on the actual current price.

Exercise 9B

1. Adapt Fig. 9.3 to show that period $t + 1$ profit is larger if the firm's expectation of period $t + 1$ price is correct, than if it is incorrect.
2. Will the firm have a larger profit if its expectation of p'_f is 10 per cent larger than $p'_a{}^{t+1}$, or 10 per cent smaller?
- 3.* Analyse the relationship between the short- and long-run decisions if the actual and forecast output prices are equal and constant, but the firm's forecast of the price of its variable input may differ from its actual price.
- 4.* Assume that the firm has correctly forecast current price and believes that next period's price will be the same as this period's. Suppose that this forecast is incorrect and the actual price in period $t + 1$ is less than forecast, but that the firm correctly forecasts that the price in period $t + 2$ will remain at the actual level for period $t + 1$. Show that for *small* changes in the actual price the long-run response exceeds the short-run, i.e. that the long-run supply curve is *more* elastic than the short-run. Draw SMC and LMC curves and the corresponding total cost curves which will lead to the long-run supply elasticity being (a) more and (b) *less* than the short-run for *large* price changes.

C. The multi-product firm*

In this and the next section we will not use the two-stage optimization procedure (deriving a cost function and then maximizing the difference between revenue and cost) of the previous sections. We will instead adopt the single-stage procedure of simultaneous choice

input and output levels. These two sections are also concerned with the firm's *long-run* decision or plan, it being assumed that there are no constraints on the adjustment of inputs. It is also assumed that the actual and forecast price are always equal and constant.

The notation of section 7D will be adopted in this section, so that the firm's decision variables are its *net* output levels $y = (y_1, \dots, y_n)$. Recall from that section that if $y_i < 0$ good i is an input and so $p_i y_i$ will be negative and measure the outlay on good i by the firm. If $y_i > 0$ good i is an output so that $p_i y_i > 0$ is the revenue from the sale of i . Since profit π is the difference between revenues and costs the firm's profit is

$$\pi = \sum p_i y_i = p y$$

The reader should also recall from section 7D that the firm's technologically feasible net output bundles can be described either by means of the implicit production function $g(y) \leq 0$, or the concept of the production set (PS). If all goods are divisible profit π will be a continuous function of the firm's net outputs and, if the PS is assumed to be non-empty, closed and bounded, the Existence Theorem of section 2B applies. (What further theorems of Chapter 2 can be applied if the PS is also strictly convex?)

When all prices are positive the firm will never choose a bundle y where $g(y) < 0$. (Readers should apply the argument of section A for the single output, two-input case to convince themselves of this.) Hence the firm's decision problem is

$$\max_y \pi = p y \text{ s.t. } g(y) = 0 \quad [C.1]$$

There are of course no non-negativity constraints on the y_i because the notational convention of this section gives an economically sensible interpretation to the negative net output levels as inputs, or goods demanded by the firm. The Lagrange function for [C.1] is $\pi + \lambda g(y)$ and the first-order conditions are

$$p_i + \lambda g_i = 0 \quad (i = 1, \dots, n) \quad [C.2]$$

$$g(y) = 0$$

Rearranging the condition on good i gives $p_i = -\lambda g_i$ and dividing by the similarly rearranged condition on good j gives

$$\frac{p_i}{p_j} = \frac{g_i}{g_j} \quad (i = 1, \dots, n; i \neq j) \quad [C.3]$$

This general condition succinctly summarizes a number of familiar results for the three logically possible cases:

1. Both goods i and j are inputs. In this case g_i/g_j is the marginal rate of technical substitution between two inputs (see section 7D) and for profit maximization this must be equated to the ratio of the inputs' prices. This is the same condition as that required for cost minimization in section 8B, which is to be expected since cost minimization is a necessary condition for profit maximization.

2. When i is an input and j an output g_i/g_j is the marginal product of i in the production of good j : MP^j_i . Rearranging [C.3] yields

$$p_j = \frac{p_i}{MP^j_i}$$

But p_i/MP_i^j is the marginal cost of good j (see section 8B) so that [C.3] states that for profit maximization the output of a good should be set at the level at which its marginal cost is equal to its price, thus confirming the results of section A. This is illustrated in Fig. 9.4 where y_1 is the firm's sole input and y_2 its sole output. The shaded area is the PS , π_1 is an iso-profit line satisfying the equation $\pi_1 = p_1 y_1 + p_2 y_2$ or $y_2 = (\pi_1 - p_1 y_1)/p_2$, and π_2, π_3 are derived in a similar way. The profit maximizing net output bundle is $y^* = (y_1^*, y_2^*)$ where the highest attainable iso-profit line, π_3 , is tangent to the upper boundary of the firm's PS . The negative of the slope of the iso-profit line is p_1/p_2 and the negative of the slope of the boundary of the PS is the rate at which y_2 increases as y_1 decreases (the input 1 is increased) or the marginal product of the input 1 in production of output 2: MP_1^2 . Hence condition 3 is satisfied at y^* : $p_1/p_2 = MP_1^2$, or $p_2 = p_1/MP_1^2$, so that price is equated to marginal cost. The firm's profit is $\pi_3 = p_1 y_1^* + p_2 y_2^* = p y^*$ or, measured in terms of the output y_2 by the intercept of the iso-profit curve on the y_2 axis: π_3/p_2 .

3. If both i and j are outputs g_i/g_j is the marginal rate of transformation between them (MRT_{ij}), so that when the firm produces more than one output it will maximize profit by producing where the MRT between two outputs is equal to the ratio of their prices. This is illustrated in Fig. 9.5 where the firm produces two outputs (y_1, y_2) from the single input, good 3. The three transformation curves show the varying combinations of the outputs that can be produced from different fixed input levels. The $g(y_1, y_2, y_3^1) = 0$ curve, for example, shows the combinations of goods 1 and 2 that can be produced in a technically efficient way when good 3 (the input) is fixed at y_3^1 . As the input level is increased the transformation curve shifts outward.

R_1, R_2 and R_3 can be called *iso-revenue lines*. They show output bundles which will produce the same total revenue: $p_1 y_1 + p_2 y_2 = R_j, j = 1, 2, 3$ where R_j is a given constant, $R_1 < R_2 < R_3$. Thus the lines have the equation $y_2 = (R_j - p_1 y_1)/p_2, j = 1, 2, 3$.

For a given level of the input y_3 , the firm's costs are given and so it maximizes profit by choosing an output combination which maximizes revenue. If, for example, $y_3 = y_3^1$ the firm will choose the output bundle y^1 , where the highest attainable iso-revenue line R_1 is tangent to the transformation curve generated by $y_3 = y_3^1$. If $y_3 = y_3^2$ or $y_3 = y_3^3$ the firm

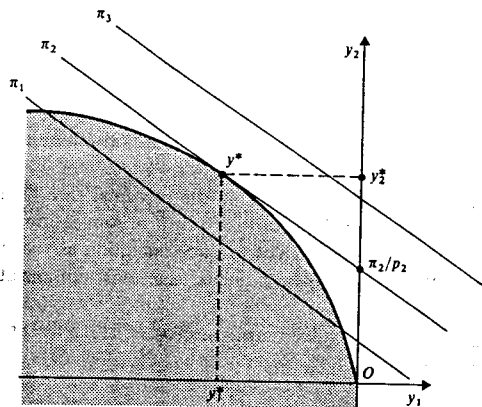


Fig. 9.4

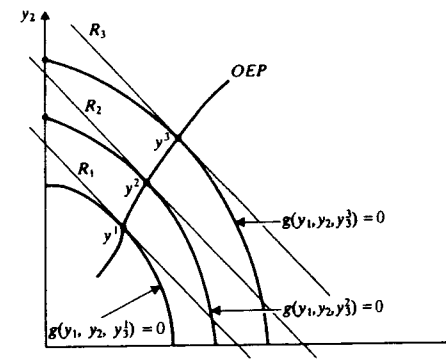


Fig. 9.5

chooses y^2 or y^3 where the respective transformation curves are tangent to iso-revenue lines. OEP is the *output expansion path*: the locus of points such as y^1, y^2, y^3 , generated by the transformation curve shifting as the input level varies. The firm's profit varies as it moves out along OEP since its revenue is increasing (higher iso-revenue curves are reached) and so is its cost (larger inputs are required to reach higher transformation curves). The firm will choose the points on OEP where the difference between revenue and cost is at a maximum. If for example this is y^2 , the firm's profit is $p_1 y_1^2 + p_2 y_2^2 + p_3 y_3^2 = R_2 + p_3 y_3^2$ (where of course y_3^2 is negative since good 3 is an input).

Exercise 9C

- 1.* Suppose that a multi-product firm is decentralized into autonomous product divisions, where each product is sold in a competitive market.
 - (a) Show that if the production function is separable, maximization of the profit of each division will lead to maximization of the profit of the firm as a whole.
 - (b) Conversely, demonstrate that separate profit maximization is not optimal if the production function is not separable with the joint cost being charged to the different divisions in proportion to (i) the price of the product; (ii) revenue from each product; (iii) the separate costs of each division.
 - (c) Should a division which makes a loss under one of the above joint cost allocations be closed down?
 - (d) If not, under what circumstances should a division be closed down?

D. The profit function and comparative statics*

Using the terminology of section 8D, the net output or netput vector y^* which solves the firm's profit maximization problem [C.1] is a function of the price vector p : $y^* = y(p)$

and hence so is the firm's maximized profit

$$\Pi = py^* = py(p) = \sum_i p_i y_i(p) = \Pi(p) \quad [\text{D.1}]$$

The profit function $\Pi(p)$ has a number of properties which are useful in deriving predictions about the firm's response to price changes.

(a) $\Pi(p)$ is increasing in p_i if the firm supplies good i ($y_i(p) > 0$) and decreasing in p_i if good i is an input used by the firm ($y_i(p) < 0$). We leave it to the reader to prove that firms are made better off by increases in the prices of goods that they sell and worse off by increases in the prices of goods that they buy. (Use the Envelope Theorem).

(b) $\Pi(p)$ is linear homogeneous in p . This follows from the fact that if $y(p)$ is profit maximizing at prices p then it is feasible [$y(p) \in PS$] and satisfies

$$py(p) \geq py \quad \text{all } y \in PS \quad [\text{D.2}]$$

which implies

$$tpy(p) \geq tpy \quad \text{all } y \in PS, t > 0 \quad [\text{D.3}]$$

Thus $y(p)$ is also optimal at prices tp and profit at prices tp is

$$\Pi(tp) = tpy(tp) = tpy(p) = t\Pi(p)$$

[D.3] also implies that the firm's optimal netput bundle is unaffected by equal proportionate changes in prices: the net supply functions of the firm are homogeneous of degree 0 in prices:

$$y_i(tp) = y_i(p) \quad [\text{D.4}]$$

(c) $\Pi(p)$ is convex in p . The proof is similar to that used to establish the concavity of the consumer's cost or expenditure function in section 4A. Consider three price vectors p^0 , p^1 and $\bar{p} = tp^0 + (1-t)p^1$. Using the definition [D.2] of the netput vector which is profit maximizing at p implies

$$p^0 y(p^0) \geq p^0 y(\bar{p}) \quad \text{and} \quad p^1 y(p^1) \geq p^1 y(\bar{p})$$

which in turn implies

$$tp^0 y(p^0) \geq tp^0 y(\bar{p}) \quad \text{and} \quad (1-t)p^1 y(p^1) \geq (1-t)p^1 y(\bar{p}) \quad [\text{D.5}]$$

for $0 \leq t \leq 1$. Using the definition of the profit function [D.1], adding the left-hand side of the first inequality in [D.5] to the left-hand side in the second and similarly for the right-hand sides, gives

$$\begin{aligned} t\Pi(p^0) + (1-t)\Pi(p^1) &\geq tp^0 y(\bar{p}) + (1-t)p^1 y(\bar{p}) \\ &= [tp^0 + (1-t)p^1] y(\bar{p}) = \bar{p} y(\bar{p}) = \Pi(\bar{p}) \end{aligned} \quad [\text{D.6}]$$

which establishes the convexity of $\Pi(p)$.

(d) *Hotelling's lemma*: $\partial \Pi(p) / \partial p_i = y_i(p)$. We can prove this by adapting the argument of section 8B used to establish Shephard's lemma. Define the function

$$G(p, p^0) = \Pi(p) - py(p^0) \geq 0$$

which cannot be negative because $y(p^0)$ is profit maximizing at p^0 and cannot yield a greater profit at p than $y(p)$ which is profit maximizing at p and which yields profit

$\Pi(p) = py(p)$. Since G is minimized with respect to p at $p = p^0$ (where $G(p^0, p^0) = 0$) its partial derivatives with respect to p_i must be zero:

$$\left. \frac{\partial G(p, p^0)}{\partial p_i} \right|_{p=p^0} = \left. \frac{\partial \Pi(p)}{\partial p_i} \right|_{p=p^0} - y_i(p^0) = 0 \quad [\text{D.7}]$$

and since $\Pi_i(p^0) = y_i(p^0)$ must be true for all p^0 , Hotelling's lemma is established.

Suppose that the firm's technology ensures that the profit function is twice continuously differentiable. Hotelling's lemma can then be used to conclude that *cross-price effects on net supply functions are equal*. A function which is twice continuously differentiable has equal cross-partial derivatives so $\Pi_{ij}(p) = \Pi_{ji}(p)$ and using [D.7] we see that $\partial y_i(p) / \partial p_j = \partial y_j(p) / \partial p_i$. Convexity implies further restrictions on the second-order partials of $\Pi(p)$ and thus, using Hotelling's lemma, on the changes in the net supplies induced by price changes. In particular, convexity implies that the second-order partial derivatives of $\Pi(p)$ are non-negative and so

$$\frac{\partial y_i(p)}{\partial p_i} = \frac{\partial^2 \Pi(p)}{\partial p_i^2} \geq 0 \quad [\text{D.8}]$$

or: the firm's net supply of a good never decreases with its price.

When good i is an output this result confirms the result derived in the special case of the single-output firm in section A where the supply curve of output was positively sloped. When good i is an input an increase in its price causes the firm to use less of it (if $y_i < 0$ then an increase in y_i corresponds to a reduction in the use of an input). Thus the firm's demand curve for an input can never be positively sloped. In section 8B we used Shephard's lemma to show that the demand for an input at given output level could not be increased by an increase in its price, i.e. the own price substitution effect was non-positive. The result derived here, using Hotelling's lemma, is much more powerful because it takes account both of the substitution effect and of the fact that a change in the input price will generally change the firm's maximizing output as well (the output effect), leading to a further change in the demand for the input.

If the firm's technology implies that the profit function is not twice continuously differentiable we can still use the definition [D.2] of the profit maximizing net output vector $y(p)$ to make predictions about its response to price changes. Let p^0 and p^1 be two price vectors and $y(p^0)$ and $y(p^1)$ be the respective profit maximizing net output vectors. Then from the definition [D.2]

$$p^0 y(p^0) - p^0 y(p^1) = p^0 [y(p^0) - y(p^1)] = p^0 \Delta y \geq 0 \quad [\text{D.9}]$$

and

$$p^1 y(p^1) - p^1 y(p^0) = p^1 [y(p^1) - y(p^0)] = -p^1 \Delta y \geq 0 \quad [\text{D.10}]$$

Adding [D.9] and [D.10] gives

$$p^0 \Delta y - p^1 \Delta y = (p^0 - p^1) \Delta y = \Delta p \Delta y \geq 0 \quad [\text{D.11}]$$

which is the *fundamental inequality of profit maximization*.

[D.11] is a strong result because it requires only that the firm's profit maximization problem has a solution (not necessarily unique) for all p so that the profit function is well

defined. It is not necessary that the profit function be differentiable. [D.11] says that the sum of the product of the price changes and the net supply changes $\sum \Delta p_i \Delta y_i$ must be non-negative and can be used to test the profit maximization hypothesis. If only one price changes [D.11] reduces to

$$\Delta p_i \Delta y_i \geq 0 \quad [\text{D.12}]$$

which confirms our earlier conclusion, reached via Hotelling's lemma and the convexity of $\Pi(p)$, that increases in the price of good i do not reduce the supply of good i if it is an output and do not increase the firm's demand for it if it is an input.

Corporation taxes

Suppose that the firm must pay a percentage tax on its profits. The net of tax profit is $(1 - t)py$ where t is the percentage rate of corporation tax. (Since in the long run the firm can always earn a zero profit by ceasing production, py can be safely considered to be non-negative.) If both sides of [D.2] are multiplied by $(1 - t)$ we get

$$(1 - t)py(p) > (1 - t)py$$

If $y(p)$ maximizes pre-tax profit it will also maximize after-tax profit. Hence the rate of corporation tax will have no effect on the profit maximizing firm's net supply decisions.

The reader is warned that this result applies to a tax levied on what economists usually define as profit, namely the difference between revenue and *all opportunity costs*. Most of the taxes which are in practice called 'profits' taxes or corporation taxes, however, are taxes on the difference between revenue and the costs *allowed by the tax authorities*. If there is any difference between opportunity costs and the allowable costs a 'profits' tax may lead to a change in the firm's behaviour, as we see in Fig. 9.6.

Two kinds of divergences between opportunity and allowable costs are likely to be important. First, the funds invested by the owners in a firm will have an opportunity cost (the return which could have been earned in alternative uses of the funds) but, unlike say interest charges on bank loans to the firm, this opportunity cost is not usually counted as an allowable cost in calculating the taxable profit. Second, in a period of rapid inflation the recorded cost of inputs used by the firm, which is the allowable cost of the inputs for tax

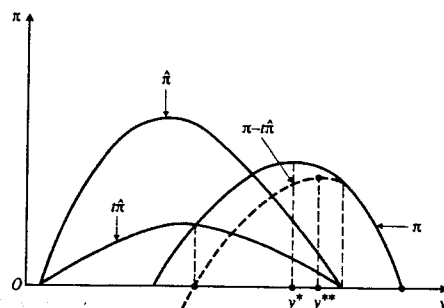


Fig. 9.6

purposes, will be less than their opportunity costs if there is any appreciable lag between purchase and use of the inputs. In either of these cases some of the opportunity costs will be disallowed for calculation of the taxable profit, which will therefore exceed the true profit.

It should also be noted that the tax authorities' definition of revenue may also differ from that of the economist and this will be a further reason why we would expect actual 'profits' taxes to alter the behaviour of firms.

Figure 9.6 illustrates our remark about the importance of the distinction between opportunity and allowable costs. $\hat{\pi}$ plots taxable profit, which differs from π because some opportunity costs are not recorded or are disallowed. If a tax is levied on taxable profits $\hat{\pi}$ the firm's tax bill is $t\hat{\pi}$, which is also plotted in Fig. 9.6. The firm's after-tax pure profit, which it wishes to maximize, is $(\pi - t\hat{\pi})$ which is drawn as the dashed line. The tax on taxable profit will therefore alter the firm's output from y^* , which maximizes before-tax pure profit, to y^{**} , which maximizes after-tax pure profit $(\pi - t\hat{\pi})$. Similarly, changes in t will change the $(\pi - t\hat{\pi})$ -maximizing output.

Lump-sum taxes and fixed costs

Let T be some lump sum tax or fixed cost that the firm must pay whatever its output level. Then the firm's net profit is $py - T$ and if T is subtracted from both sides of [D.2] we have

$$py(p) - T > py - T$$

and if $y(p)$ maximizes profit before tax or the fixed cost it will maximize net profit after the tax or fixed cost. The level of lump-sum taxes will have no effect on the firm's decisions. This result is critically dependent on T being independent of py . Suppose, for example, that the firm had to pay a licence fee in order to operate. This licence fee is *not* a lump sum tax or fixed cost because if the firm does not operate, i.e. $y = (0, \dots, 0)$ it does not have to pay the fee. The fee therefore varies discontinuously with the firm's net output decision. If, for example, the firm's optimal bundle is $y^* > (0, \dots, 0)$ raising T from $0 < T < py^*$ to $T > py^*$ will cause the firm to switch from y^* to $(0, \dots, 0)$: it will go out of business.

These results are illustrated in Fig. 9.7 for a single-product firm. π is the non-negative part of the pre-tax profit curve in the figure. Pre-tax profit is maximized at y^* . A proportional

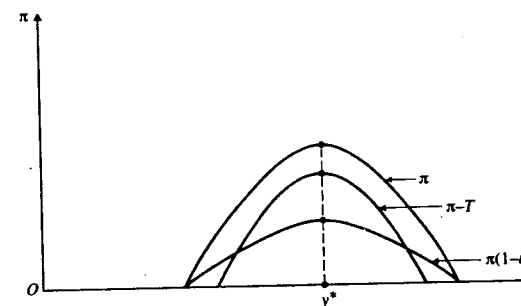


Fig. 9.7

profit tax of t will give rise to the after-tax profit curve which plots $(1-t)\pi$. The proportional tax flattens the profit curve, as (except where $d\pi/dy = 0$) the slope of π is $(1-t)d\pi/dy < d\pi/dy$. y^* also maximizes after-tax profit and so changes in t do not affect the output at which profit is maximized. Lump-sum taxes T shift the profit curve vertically downward to $\pi - T$. Again, no change in after-tax profit maximizing output is caused by changes in T , as long as optimal output remains positive, by changes in T .

Exercise 9D

1. Draw diagrams of the firm's PS in the single input, single output case to illustrate the circumstances under which
 - (a) the profit maximizing y is not unique,
 - (b) the same y is optimal at different relative prices.

What do the firm's net supply curves look like in these two cases?

2. Suppose that in the time between the firm's purchase and use of inputs *all* prices (including the price of its output) double. How, if at all, will recorded profit (revenue minus the purchase cost of the inputs) differ from actual or pure profit? Does a percentage tax on recorded profit lead to a rise or fall in the firm's output? (Assume that the firm realizes that the tax is levied on recorded profit and it correctly anticipates the rate of inflation.)

References and further reading

There is a comprehensive analysis of profit functions and their relationship with cost functions in

R. G. Chambers. *Applied Production Analysis: A Dual Approach*, Cambridge University Press, Cambridge, 1988, ch. 4,

and at a more advanced level in

D. McFadden. 'Cost, revenue and profit functions', in M. Fuss and D. McFadden (eds), *Production Economics: A Dual Approach to Theory and Applications*, North Holland, Amsterdam, 1978, vol. 1.

The classic reference on cost and supply curves is

J. Viner. 'Cost curves and supply curves', in G. J. Stigler and K. E. Boulding (eds), *Readings in Price Theory*, George Allen and Unwin, London, 1953.

Various concepts of income and profit are examined in

R. H. Parker and G. C. Harcourt. *Readings in the Concept and Measurement of Income*, Cambridge University Press, London, 1969, ch. 7.

CHAPTER 10

The theory of competitive markets

In the preceding seven chapters we have considered models of the optimal choices of consumers and firms. In these models, prices were always taken as parameters outside the control of the individual decision-taker. We now examine how these prices are determined by the interaction of the decisions of such 'price-taking' individuals. Since this interaction takes place through markets, we examine theories of the operation of markets whose participants act as price-takers, that is, of *competitive markets*. In later chapters we examine markets in which some of the decision-makers perceive that they are able to influence the price, which therefore becomes a choice variable in their decision problem.

In Chapters 8 and 9 we drew a distinction between production and supply in the short run and in the long run. We maintain that distinction in market analysis, since supply conditions are an important determinant of the market outcome. We again think of demand and supply as rates of flow per unit time. The short run is the period over which firms have fixed capacity. In the long run all inputs are variable. For example, if it takes a year to plan and implement capacity changes then the short run is this year and the long run is next year. Since decisions for the long run are necessarily *planning* decisions, expectations must come into the picture. Realistically, so should uncertainty, but we postpone consideration of this to later chapters of the book.

This chapter adopts a *partial equilibrium* approach: a single market is considered in isolation. This is not entirely satisfactory, since there are strong interactions between markets. For example, we shall see that in aggregating firms' supply curves to obtain a market supply curve we should take account of the effect of expansion of aggregate market output on the prices of inputs used by the firms. We need a general equilibrium analysis in which market interactions are fully taken into account. This is provided in Chapter 16. The justification for a partial equilibrium analysis is that it is simple and can give useful insights. Moreover, the key issues concerning the existence and stability of equilibrium – the central concepts with which we shall be concerned – can be considered in a particularly simple context.

A. Short-run equilibrium

Let $x_i = D_i(p)$ be the i th consumer's demand for the commodity at price p and

$$x = \sum_i x_i = \sum_i D_i(p) = D(p) \quad [\text{A.1}]$$

be the market demand function. The short-run supply function of firm j is

$$y_j = s_j(p, w) \quad [\text{A.2}]$$

where y_j is the output of firm j and w is the price of the variable input.

It might appear that we could proceed to obtain a market supply function by aggregating the firms' supply functions as we did the consumers' demand functions in [A.1], but this is not in general the case. In deriving the firm's supply function in Chapter 8 we assumed input prices constant. This was a natural assumption to make, since any one firm in a competitive 'industry' (defined as the set of all producers of a given commodity) could be expected to be faced with perfectly elastic input supply curves. Then, as its output price is raised, the firm could expand its desired production and input levels without raising input prices. This assumption may not be appropriate for the industry as a whole, however: as the price at which they can sell their outputs rises for all firms, expansion in production and input demands may raise input prices because the size of demand increase is no longer insignificant, and input supply functions have positive slopes to the industry as a whole.

Thus denoting the total amount of the variable input used by the industry by $z(y)$ ($z'(y) > 0$), if we have

$$w = w(z(y)) \quad [\text{A.3}]$$

with $w'(z) > 0$, there are *pecuniary external diseconomies*: an increase in the total output of firms in the industry increases the price of an input.

The consequences for the firm's actual supply are shown in Fig. 10.1. In the figures, price is assumed to rise from p to p' . The firm's initial supply ($\equiv SMC$) curve is in each

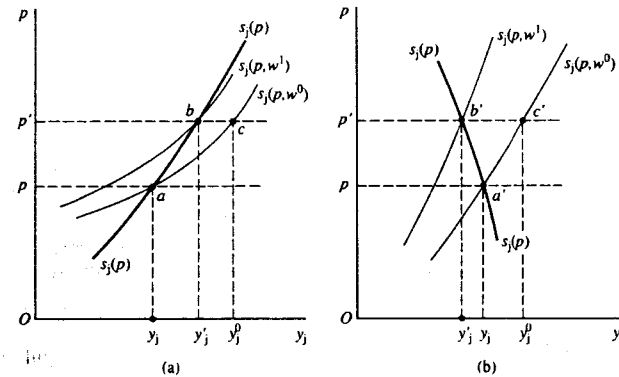


Fig. 10.1

case $s_j(p, w^0)$. However, if the effect of simultaneous expansion by all firms is to raise input prices from w^0 to w^1 the marginal cost curves and short-run supply curves of each firm must rise. Figure 10.1(a) shows one possible result of the expansion of firms in response to the higher price. The short-run supply curve has risen to $s_j(p, w^1)$ and so at price p' the firm will want to supply y_j' and not y_j'' . Hence the points on the firm's supply curve corresponding to p and p' , when all firms expand, are a and b respectively and $s_j(p)$ is the locus of all such price-supply pairs. Clearly, the firm's *effective market supply curve* $s_j(p)$ will be less elastic than its *ceteris paribus* supply curve $s_j(p, w)$. They would coincide if input prices were not bid up by simultaneous expansion of output by all firms (and there were no technological externalities – see question 2).

In (b) of the figure is shown a more extreme case. The increase in input prices causes a sufficient shift in the firm's *SMC* curve to make the post-adjustment output y_j' actually less than y_j , and so its effective market supply curve s_j has a negative slope. Thus, although the 'law of diminishing returns' ensures that each firm's *ceteris paribus* supply curve has a positive slope this is not sufficient to ensure that the firm's effective supply curve has a positive slope, if input prices increase with the expansion of outputs of all firms.

Denoting the effective industry supply function by $y(p)$ and substituting [A.3] in [A.2] gives the effective supply function of firm j :

$$y_j = s_j(p, w(z(y))) = s_j(p) \quad [\text{A.4}]$$

and summing gives the effective industry supply function

$$y = \sum_j y_j = \sum_j s_j(p) = s(p) \quad [\text{A.5}]$$

Differentiating [A.4] with respect to the market price gives the effective supply response of firm j (after allowing for the effect of the increase in w induced by the change in output of all firms) as

$$\frac{dy_j}{dp} = s_{jp} + s_{jw}w'(z)z'(y) \frac{dy}{dp} = s_j'(p) \geq 0 \quad [\text{A.6}]$$

Since $s_{jp} = \partial s_j(p, w)/\partial p > 0$ and $s_{jw} = \partial s_j/\partial w < 0$ we see that the firm's effective supply could be increasing or decreasing in p .

The change in industry supply as a result of the increases in p is the sum of the effective changes in the firms' supplies and so from [A.5]:

$$\frac{dy}{dp} = \sum_j \frac{dy_j}{dp} = \sum_j s_{jp} + w'z' \frac{dy}{dp} \sum_j s_{jw} \quad [\text{A.7}]$$

Solving for dy/dp gives

$$\frac{dy}{dp} = \frac{\sum_j s_{jp}}{1 - w'z' \sum_j s_{jw}} > 0 \quad [\text{A.8}]$$

Thus the effective industry supply curve is positively sloped despite the fact that some of the firms may have negatively sloped effective supply curves. The slope of this market supply function depends on the extent to which increases in input demands increase input prices and the consequent increases in marginal costs at all output levels. Note that at a

market supply $s^0 = s(p^0)$, i.e. a point on this supply function, each firm's marginal cost is exactly equal to p^0 , given that all output adjustments have been completed. We define p^0 as the *supply price* of the corresponding rates of output y_j^0 since it is the price at which each firm would be content to supply – and to go on supplying – the output y_j^0 . At any greater price firms would find it profitable to expand production; at any lower price, they would wish to contract.

Figure 10.2 shows a number of possible situations which might arise when we put the supply function together with the demand function. In (a) we show a 'well-behaved' case. The price p^* , with desired demand x^* equal to supply y^* is obviously an equilibrium, since sellers are receiving the price they require for the output they are producing, and this output is being taken off the market by buyers at that price. There is no reason either for sellers to change their output (since each $y_i^* = s_i(p^*)$ maximizes j 's profit at price p^*) or for buyers to change the amount they buy.

Figure 10.2(b) represents a case which could arise when there is a certain kind of discontinuity in the supply curve $s(p)$. Recall from Chapter 9 that when price falls below average variable cost (AVC) a profit maximizing firm will produce zero output. If all firms have identical AVC s, they will all produce zero output at the same price. Hence, at some critical price, shown as p^0 in the figure, supply may suddenly drop to zero. Thus there is a discontinuity in the short-run supply function at p^0 . If it happens that the demand curve has the position shown, there is no equilibrium. Individual buyers would be prepared to offer individual sellers prices in excess of p^0 for some output but if firms respond by starting up production, they flood the market and price must fall to a level below p^0 . Discontinuities, which are perfectly possible in this case, cause problems for the existence of equilibrium. Note, however, that continuity is sufficient but not necessary: if $D(p)$ were higher and intersected $s(p)$, as in (a) of the figure, the discontinuity at p^0 would present no difficulty.

In (c) we show a further possibility. Suppose that firms do not all have the same AVC , but instead are evenly distributed over a range of AVC s, with the minimum point of the lowest AVC curve being equal to p'' . If there are many sellers, and each seller is an insignificant part of the market, we can then take the $s(p)$ curve as continuous, with intercept at p'' . However, at price $p' < p''$, demand is zero – no one would be prepared to pay p' or more for this good. It follows that equilibrium in this market implies a zero

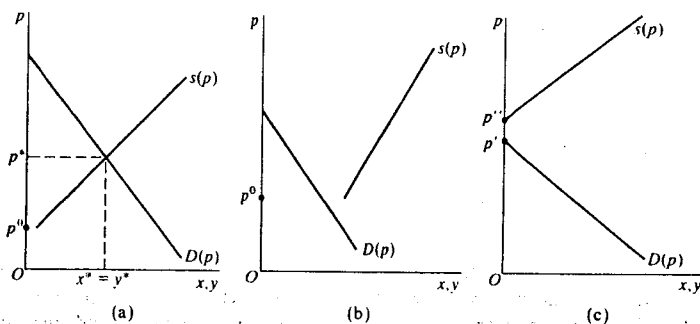


Fig. 10.2

output and a price in the interval $[p', p'']$ – the highest price any buyer would pay is insufficient to cover the AVC of the firm with the lowest minimum AVC . Thus the good is a 'non-produced good', a good for which the technology is known, which firms would supply if the price were right, but nobody wants to buy. The reader will find it instructive at this point to construct the *excess demand functions*, defined by

$$z(p) = D(p) - s(p) \quad [A.9]$$

in these three cases, and illustrate them in a price-excess demand graph of the type shown in Fig. 10.3 below.

Figure 10.2(b) suggests that the presence of a discontinuity in a supply or demand function – briefly, in the excess demand function – may cause a failure of equilibrium to exist. This is a matter of some concern, since our theory of the market predicts the market outcome to be the *equilibrium* outcome, and raises the question: what do we have to assume to *ensure* that the market has an equilibrium? In Chapter 16, we consider this question for the entire system of markets taken together. To take the case of only one market is to give only a provisional answer to the question since we then ignore the interdependence among markets. Nevertheless, it is instructive to consider the existence question in the simple context of one market.

Figure 10.2 shows that discontinuity is a problem. Is it then enough to assume that $z(p)$ is a continuous function of p ? Clearly not. An equilibrium is a price $p^* > 0$ such that $z(p^*) = 0$. If $z(p) < 0$, or $z(p) > 0$, for all $p > 0$, then $z(p)$ may be continuous but we will not have an equilibrium. This suggests the following existence theorem for a single market. If

- (a) the excess demand function $z(p)$ is continuous for $p \geq 0$;
- (b) there exists a price $p^0 > 0$ such that $z(p^0) > 0$;
- (c) there exists a price $p^1 > 0$ such that $z(p^1) < 0$

then there exists a price $p^* > 0$ such that $z(p^*) = 0$, i.e. an equilibrium price.

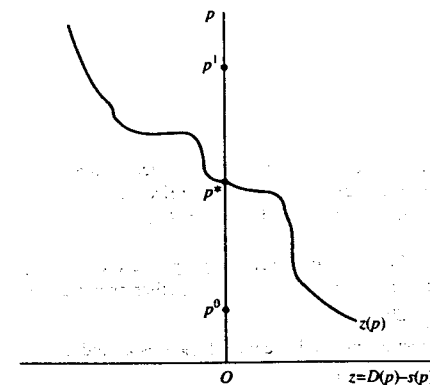


Fig. 10.3

The intuition is clear from Fig. 10.3. If the excess demand curve is continuous and passes from a point at which excess demand is positive to a point at which excess demand is negative, it *must* cross the price axis, giving an equilibrium price.

The significance of the equilibrium price is that it induces buyers to demand exactly the output that results from individual sellers' profit maximizing decisions at that price. Plans are all mutually consistent and can be realized. If equilibrium exists, it is the solution to the resource allocation problem for the market in question. We now turn to the equally important question of the stability of a market in the short-run.

Exercise 10A

1. *External pecuniary economies*. Derive the market and firm effective supply functions on the assumption that input prices fall as all firms expand output. What could account for this?
2. *Technological external diseconomies* exist when an increase in industry output increases all firms' costs. For example, owners of oil wells drilled into the same oil field may find that it is more expensive to produce any given output from their well when total output from the field is larger. Congestion of fishing grounds is another example. Suppose that as the industry output expands all firms' short-run *marginal* costs are increased. Apply the analysis of pecuniary diseconomies to show that the effective supply of some firms may be negatively sloped but the effective industry supply curve will be positively sloped. (*Hint*: write the supply function of firm j as $y_j = s_j(p, a)$ where $a(y)$, $a' > 0$ is a shift parameter reflecting the external diseconomy: $s_{ja} < 0$.)
3. Suppose that (a) the market supply function $s(p)$ is continuous and non-decreasing in p with $s(p) > 0$ for $p > p^0$; (b) the demand function $D(p)$ is continuous, non-increasing in p , with $D(p) \rightarrow 0$ as $p \rightarrow \infty$ (because consumers have finite incomes) and $D(p^0) > 0$. Are these assumptions sufficient to ensure the existence of an equilibrium in the market?
4. The supply curve of labour $s(w)$ may be backward bending for some range of wage rates (recall section 4C). Does this mean that there may be no equilibrium in the market for labour even if the labour demand curve $D(w)$ is continuous and strictly decreasing in w ?
5. *Incidence of taxes*. Consider a market in which a per unit tax t is levied so that $p_s = p_c - t$, where p_s is the price received by suppliers and p_c the price paid by consumers. The supply function is $s(p_s)$ and the demand function $D(p_c)$.
 - (a) Show that the *economic incidence* of the tax (its effects on p_s , p_c and the quantity traded) are independent of the *legal incidence*, i.e. whether producers or consumers must pay the tax to the government.
 - (b) Show that legal incidence does affect economic incidence if there is a binding maximum price in the market.
 - (c) What happens if there is a binding minimum price?

B. Stability of equilibrium

The analysis of stability is concerned with the question: if at a given point in time the market price is not at an equilibrium value, will changes take place over time which cause the price to converge to such a value? If the answer is yes, the market is stable, and if no, unstable. There is a related aspect of stability analysis, which is concerned with the properties of a *particular* equilibrium price. We can ask whether, if price is not equal to that *particular* value, it will tend towards it, and if so, we call that equilibrium stable, and if not, unstable. Clearly, the two aspects of stability boil down to the same thing if there is a unique equilibrium in a market. However, where there are multiple equilibria, a particular equilibrium price may be unstable but the market may be stable, as we shall see. In general, we are interested in the stability of the market, rather than of a specific equilibrium position. A further point of definition: a market is *locally* stable if it tends to an equilibrium when it starts off in a small neighbourhood of one, and it is *globally* stable if it tends to an equilibrium *wherever* it starts off. Global stability implies local stability but not conversely. In general, we are more interested in global stability.

Formally a market is stable if p^* is an equilibrium price and

$$\lim_{t \rightarrow \infty} p(t) = p^*$$

where $t \geq 0$ is time, $p(t)$ is the time path of price and the initial price $p(0) \neq p^*$ is given.

The analysis of stability is concerned with a market's *disequilibrium* behaviour. It follows that we have to formulate a theory of how markets operate out of equilibrium. Any such theory must, implicitly or, preferably, explicitly, assume answers to three fundamental questions:

1. How do the market price or prices respond to non-zero excess demand?
2. How do buyers and sellers obtain information on the price or prices being offered and asked in the market?
3. At what point does trading actually take place, i.e. when do buyers and sellers enter into binding contracts?

These questions are important because answers to them may differ and differences in the answers lead to significant differences in the models of disequilibrium adjustment to which the theories give rise. In questions 1 and 2 we use the phrase 'price or prices' because at this stage we prefer to keep our options open. Some theories may provide for a single price to prevail throughout the market even out of equilibrium, while others allow there to be differences in prices offered by buyers and asked by sellers throughout the market. In fact, whether or not a unique price will always prevail depends very much on the answers adopted to questions 2 and 3.

To begin with we consider two continuous time models of market adjustment. The first, known as the *tâtonnement process* (tâtonnement can be interpreted as 'groping') was proposed by L. Walras. The second, which it can be argued is better suited to markets with production, was suggested by A. Marshall.

The tâtonnement process (TP)

The TP is an idealized model of how a market may operate out of equilibrium, in the sense that it may not describe the way a market works, but under certain conditions a real market may operate as if its adjustment process were a TP. There is a central individual who can be called the market 'umpire', and who essentially has the role of a market coordinator. He announces to all decision-takers a single market price (the answer to Question 2), which they take as a parameter in choosing their planned supplies or demands. They each inform the umpire of their choices and he aggregates them to find the excess demand at the announcement price. He then revises the announced price by the following rule (the answer to Question 1):

$$\frac{dp}{dt} = \lambda z(p(t)) \quad \lambda > 0 \quad [\text{B.1}]$$

that is, he changes the price at a rate proportionate to the excess demand. No trading takes place unless and until equilibrium is reached (the answer to Question 3) at which time sellers deliver their planned supply and buyers take their planned demand. Notice that in this process there need be no contact between a buyer and seller – everything is mediated through the umpire.

Figure 10.4 shows three possible cases of market excess demand functions. In (a), the excess demand curve has a negative slope. It follows that if initially the umpire announces the price $p^0 < p^*$, excess demand will be positive and he will revise the announced price upward towards p^* ; if the announced price were above p^* it would be revised downwards. Since these movements are always in the equilibrating direction, from wherever the process starts, equilibrium will be globally stable.

In (b), the excess demand curve has a positive slope. If the announced price is initially at p^0 the umpire will now reduce price, since $z < 0$, and hence the TP leads away from equilibrium. A similar result would occur if the initial price were above p^* . Hence in this market the equilibrium is globally unstable.

In (c) we have a somewhat more complex case. The excess demand curve is backward-bending, having a negative slope over one range of prices and a positive slope over another.

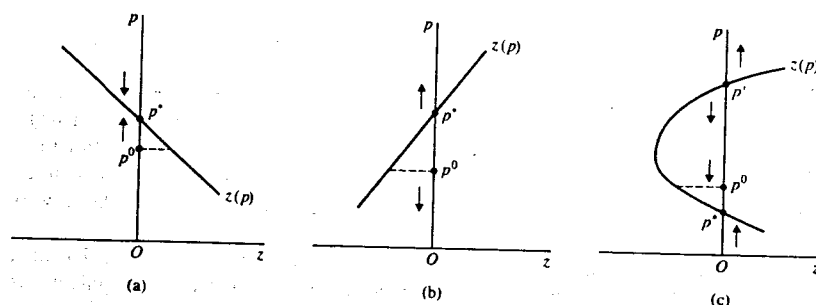


Fig. 10.4

In this case if the initial price were anywhere in the interval $0 \leq p < p'$, the TP would converge to the equilibrium p^* . If, however, the initial price were, say, $p'' > p'$, the market would move away from equilibrium, since excess demand is positive over this range and so price would be increased. Therefore the market is not globally stable, since an initial point sufficiently far from the equilibrium p^* would lead away from market equilibrium. This market has two equilibrium positions, one at p^* and one at p' ; the former is locally but not globally stable, the latter is locally (and therefore globally) unstable.

From this discussion we can deduce the following stability conditions, i.e. sufficient conditions for the TP to be stable:

- (a) equilibrium is globally stable if excess demand is positive whenever price is less than its equilibrium value and negative when price is above its equilibrium value;
- (b) equilibrium is locally stable if the condition (a) holds for prices in a small neighbourhood of an equilibrium

We can put this stability analysis more formally by using the device of a distance function, a function giving the distance between two points (see the Appendix to Chapter 16 for a fuller discussion). Thus define the distance function

$$\delta(p(t), p^*) = (p(t) - p^*)^2 \quad [\text{B.2}]$$

which measures the distance between an equilibrium price p^* and some other price $p(t)$ (note: $\delta(p(t), p^*) > 0 \Leftrightarrow p(t) \neq p^*$). A necessary condition for the time path of price $p(t)$ to converge to p^* is that $d\delta/dt < 0$, i.e. the distance between the price path and p^* is falling through time. Differentiating we have

$$\frac{d\delta}{dt} = 2(p(t) - p^*) \frac{dp}{dt} = 2(p(t) - p^*) \lambda z(p(t)) \quad [\text{B.3}]$$

from [B.1]. Then clearly $d\delta/dt < 0$ if and only if $(p(t) - p^*)$ and $z(p(t))$ have opposite signs, consistently with the above stability condition. Note also that this is true regardless of the value of λ : this 'speed of adjustment' parameter determines only how fast, and not whether, the TP converges to equilibrium.

Is the condition also sufficient for convergence, however? It may seem 'intuitively obvious' that it is, but consider the example of the function $y = a + 1/t$. Here we have $dy/dt < 0$, but $\lim_{t \rightarrow \infty} y = a$, so we have to provide a further argument to justify the claim that $\delta(p(t), p^*)$ is not bounded away from zero.

We do this by establishing a contradiction. Suppose, without loss of generality, that $p(0) > p^*$, and suppose that $\lim_{t \rightarrow \infty} p(t) = \bar{p}$ where $\bar{p} > p^*$. The interval $[p(0), \bar{p}]$ is non-empty, closed and bounded and the function $d\delta/dt$ is continuous, so at some t we must have that $d\delta/dt$ takes on a maximum, by Weierstrass' Theorem. Call this maximum s^* . Note that since for $p(t) \neq p^*$, we must have $d\delta/dt < 0$, then $s^* < 0$ also. For any arbitrary $t = \bar{t}$, integrate to obtain:

$$\int_0^{\bar{t}} \frac{d\delta}{dt} dt = \delta(p(\bar{t}), p^*) - \delta(p(0), p^*) \quad [\text{B.4}]$$

and

$$\int_0^{\bar{t}} s^* dt = s^* \bar{t}$$

Then by definition of s^* we must have

$$\delta(p(\bar{t}), p^*) - \delta(p(0), p^*) \leq s^* \bar{t}$$

or

$$\delta(p(\bar{t}), p^*) \leq s^* \bar{t} + \delta(p(0), p^*)$$

By choosing \bar{t} large enough, we can make the right hand side of [B.7] negative, implying we must have on the left-hand side a negative value of the distance function, which is impossible. Thus we have the contradiction.

This proof makes precise the intuition that if $p(t)$ is always moving closer to p^* whenever $p(t) \neq p^*$, it cannot tend to anything other than p^* .

Marshall's process

Marshall suggested the following alternative to Walras' *TP*. Suppose that when sellers bring their output to market they sell it for whatever it will fetch. The demand and supply curves are constant over time. (Refer to Fig. 10.5.) If supply is less than the *equilibrium supply* y^* the price buyers will be prepared to pay if it is auctioned off to the highest bidders, the *demand price*, p_D , exceeds the *supply price*, p_s . Conversely, if supply exceeds equilibrium supply, auctioning off the available supply causes demand price to fall below supply price. Marshall argued that when demand price p_D exceeds supply price p_s sellers will expand supply, and conversely when p_D is less than p_s . This is because p_s equals each seller's marginal cost, and so $p_D > p_s$ implies output expansion increases profits, while when $p_D < p_s$ profits are increased by an output contraction. This suggests the adjustment rule:

$$\frac{dy}{dt} = \lambda(p_D(y) - p_s(y)) \quad [\text{B.8}]$$

where $p_D(y)$ is the *inverse demand function*, giving demand price as a function of quantity supplied (= quantity traded at any t) and similarly, $p_s(y)$ is the *inverse supply function* (derived from the firm's marginal cost functions as before). Note that at equilibrium quantity y^* , $p_D = p_s = p^*$.

Under what conditions is Marshall's process stable? If output expands when $p_D > p_s$ and contracts when $p_D < p_s$, then Fig. 10.5(a) suggests that when the supply and demand curves have the usual slopes, the market is stable. Figure 10.5(b) and (c) show that when the supply curve has a negative slope, the process is stable if the demand curve cuts the supply curve from above but *unstable* in the converse case. This is interesting, not only because backward bending supply curves are quite possible (recall section 4D), but also because the Walrasian *TP* has precisely the opposite outcomes in these cases: in Fig. 10.5(b), the corresponding excess demand function $z(p) = D(p) - s(p)$ increases with price and so the Walrasian *TP* would be unstable, whereas in Fig. 10.5(c), $z(p)$ has a

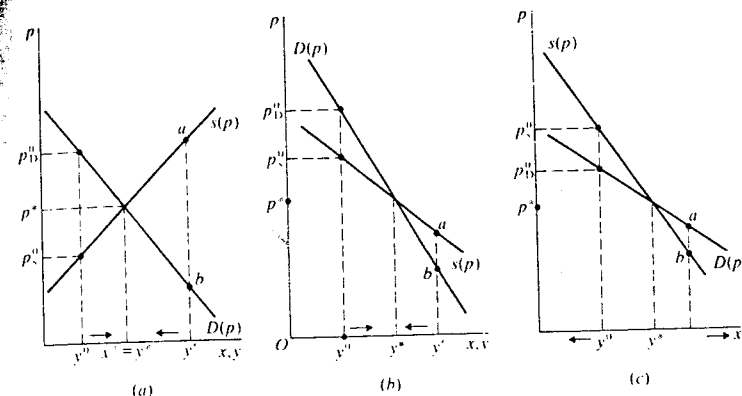


Fig. 10.5

negative slope and so the Walrasian *TP* is stable. Thus although the two adjustment processes have the same outcomes in the 'standard case', it matters which we adopt in a 'non-standard' case.

To make the stability conditions for Marshall's process more precise, we again adopt a distance function approach. Define the distance function

$$\delta(y(t), y^*) = (y(t) - y^*)^2 \quad [\text{B.9}]$$

Then

$$\frac{d\delta}{dt} = 2(y(t) - y^*) \frac{dy}{dt} = 2(y(t) - y^*) \lambda [p_D(t) - p_s(t)] \quad [\text{B.10}]$$

using [B.8]. Then, for $d\delta/dt < 0$, we require $(y(t) - y^*)$ and $(p_D(t) - p_s(t))$ to have opposite signs, confirming the diagrammatic analysis. We can establish the sufficiency of this condition along similar lines to those used in the case of the *TP* process.

We have already noted that in 'non-standard' cases the Walrasian *TP* and Marshall's process have opposite implications for market stability – it matters whether we take price as adjusting to a difference in quantities, or quantity as adjusting to a difference in demand and supply prices. We can also compare the processes in terms of the answers they give to the three questions set out at the beginning of this section:

1. *Responsiveness of price to non-zero excess demand*: in the standard case of negatively sloped demand and positively sloped supply, both processes result in market price rising (falling) when there is positive (negative) excess demand. In the Walrasian case this happens directly through the *TP*; in the Marshallian case, via the auction mechanism which establishes the demand price.
2. *Information on price(s)*: in the *TP*, this is transmitted simultaneously to all buyers and sellers by the umpire; in Marshall's process, at each instant the auction mechanism rations off available output and the demand price is immediately made known. Buyers

never need to know the supply price – sellers know their own marginal costs and so once the demand price is known an output change can result.

3. *When does trade take place?* In the *TP*, only at equilibrium; under Marshall's process, at every instant as available supply is auctioned off. Thus, we could say that Marshall's process involves *trading out of equilibrium*, with an *efficient rationing rule*, which says that available supply is auctioned off to the highest bidders. Alternatively, we could think of Marshall's process as consisting of a sequence of 'very short-run' or instantaneous equilibria, with a vertical supply curve at each of these equilibria, and the analysis then establishes conditions under which this sequence of instantaneous equilibria converges to a full equilibrium of supply and demand.

Which of these models is 'better' is, of course, an empirical question. We have to decide which process appears to capture more closely the way a particular market actually works. Walras' *TP* may seem rather unrealistic, in its reliance on a central 'umpire' collecting buying and selling intentions and announcing an equilibrium price, but some markets, for example markets in stocks and shares, and minerals such as gold and silver, are highly organized with brokers who may function much as a Walrasian umpire.

There are two features of both models which are unsatisfactory in the light of observations of how real markets work in many cases. First both processes are centralized: some device, the umpire or the auction mechanism, ensures that all buyers and sellers simultaneously face the same price. However, in many real markets, price formation is *decentralized*. Individual buyers meet, haggle and deal with individual sellers, and pressures of excess demand or supply exert their influence by causing sellers and buyers to bid prices up or down. If information on all the prices being offered and asked is fully and costlessly available throughout the market then this would be equivalent to a centralized adjustment process, but in reality this is often not the case. Buyers and sellers have to seek each other out to find the prices at which they are prepared to trade, and this *search process* is costly.

Second, in neither model do buyers and sellers *form expectations* and act upon them. In the *TP* this possibility is simply excluded. In Marshall's process, presumably some expectation of future price must be created by the observation of current demand price, and it is this which determines the change in supply, but this is not modelled explicitly, being subsumed in the adjustment rule [B.8]. In the rest of this section therefore we consider the explicit modelling of expectations in market adjustment processes.

Expectations and market stability

The concept of a *supply lag* is very important in understanding the adjustment process in many markets and to bring out its implications we move from a continuous to a discrete treatment of time. It appears in its simplest form in the market for an agricultural good, say, potatoes. At some point in time a farmer decides on the acreage of potatoes to plant. Ignoring problems such as pests, disease and adverse weather, this determines the amount of potatoes he will put on the market some time later. Thus supply of potatoes at time t , q_t^s , depends on a decision taken at $t - 1$, where the time period is the length of time between

planting and harvesting. We hypothesize that the acreage planted at $t - 1$ depends on the price the farmer expects to prevail at t . If all farmers behave like this then the market supply function is given by

$$y_t = s(p_t^e) \quad s' > 0 \quad [\text{B.11}]$$

where p_t^e is the (assumed identical) price at t all farmers expect at $t - 1$. It is assumed that demand adjusts to price at t , and so the demand function is as before,

$$x_t = D(p_t) \quad [\text{B.12}]$$

To analyse the market we have to specify how price expectations are formed. The *naive expectations* hypothesis says that

$$p_t^e = p_{t-1} \quad \text{all } t \quad [\text{B.13}]$$

That is, farmers simply assume that this period's price will continue to hold next period. This presents no problems if the market is in equilibrium over successive periods: this period's equilibrium price is also next period's equilibrium price, so the farmers' naive price expectation is correct and supply and demand will be consistent:

$$D(p_t) = s(p_{t-1}^e) = s(p_t^*) \quad [\text{B.14}]$$

where p_t^* is the equilibrium price. Suppose, however, that between $t - 1$ and t there has been a demand shift, so that p_{t-1}^* is not the equilibrium price at t . The analysis of the subsequent disequilibrium behaviour in the market is illustrated in Fig. 10.6.

In (a), p_0 was the old equilibrium price but between $t = 0$ and $t = 1$ demand has increased to $D(p)$. $y_1 = s(p_0)$ is the available supply at $t = 1$, and so when this is put on the market price rises to p_1 , where $D(p_1) = y_1$. Farmers then expect p_1 to be the market price at $t = 2$, so they plant their acreage accordingly and at $t = 2$ put $y_2 = s(p_1)$ on the market. This causes price to fall to p_2 , inducing at $t = 3$ supply of y_3 , and so on. We observe that the successive prices p_0, p_1, p_2, p_3 are converging on the new equilibrium, and so we conclude that the market is *stable*.

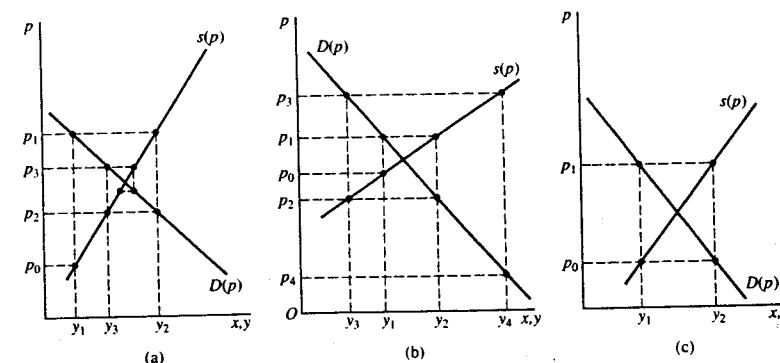


Fig. 10.6

In (b) of the figure we have a *Cobweb cycle*. The process is the same as before: price is initially p_0 , and so farmers supply $y_1 = s(p_0)$ at $t = 1$, causing price to rise to p_1 . As a result farmers plant a larger acreage and at $t = 2$ put the supply $y_2 = s(p_1)$ on the market, causing price to fall to p_2 , and so on. The interesting thing now is that the sequence of prices $p_0, p_1, p_2, p_3, \dots$ is moving away from equilibrium, therefore the market is *unstable*. In (c) we show the remaining possibility, that price moves forever between p_0 and p_1 , and so, since it never converges to equilibrium, we again have an *unstable* case.

In the figure we have drawn the supply and demand functions as linear. The difference between the stable and unstable cases is that in the stable case (a), the demand curve is less steep (referred to the quantity-axis) than the supply curve, while in the unstable case (b), the converse is true, and in (c) the slopes are exactly equal. We can put this more precisely as follows: let the supply and demand functions be, respectively

$$y_t = a + bp_{t-1}; \quad x_t = \alpha - \beta p_t \quad [\text{B.15}]$$

Since at each t we have market clearing, $x_t = y_t$, we can use [B.15] to obtain the first order linear difference equation

$$p_t = \frac{\alpha - a}{\beta} - \frac{b}{\beta} p_{t-1} \quad [\text{B.16}]$$

For stability, we require that the differences between successive prices (which are alternately positive and negative) should become successively smaller in *absolute value*. That is we require

$$-(p_{t+1} - p_t) < p_t - p_{t-1} \quad [\text{B.17}]$$

(where without loss of generality we have assumed $p_{t+1} < p_t$ and so $p_{t-1} < p_t$). Then substituting from [B.16] gives

$$\left(\frac{\alpha - a}{\beta} - \frac{b}{\beta} p_{t-1} \right) - \left(\frac{\alpha - a}{\beta} - \frac{b}{\beta} p_t \right) < p_t - p_{t-1} \quad [\text{B.18}]$$

and so

$$\frac{b}{\beta} (p_t - p_{t-1}) < p_t - p_{t-1} \quad [\text{B.19}]$$

giving the stability condition

$$b < \beta \quad [\text{B.20}]$$

Recall that b and β are the absolute values of the slopes of the supply and demand curves with respect to the *price* axes, and so [B.20] is the condition we derived from the diagram.

The naive expectations assumption is aptly named. It does seem very naive always to extrapolate the current price to the next period, even when this belief is consistently falsified. It is also unprofitable. In each period, the output which each farmer sells is not the profit maximizing output corresponding to the market price that actually prevails. To see this, note that the successive price-quantity pairs $(p_1, y_1), (p_2, y_2), \dots$, in Fig. 10.6 are not on the market supply curve, implying that each seller's marginal cost is not equal to the market price at which the corresponding output is sold. Farmers have always produced

too little when market price is high, and too much when market price is low, relative to the quantities that maximize profit at those prices. Thus, this way of forming expectations does not lead to the maximization of profits, and so, since it is behaviour inconsistent with the farmers's objective, we could call it irrational.

This reasoning led John F. Muth to propose the theory of *rational expectations*. Suppose that it is possible to estimate accurately the market demand and supply functions (for simplicity we ignore the issue of random errors of estimation). Then, it would be possible to predict the equilibrium price in each period. The person doing this (a farmer, a consultant) could make a profit by selling this information to farmers, who would find it worth buying because they can take more profitable production decisions as a result. But then, if farmers' expectations consist of the actual market outcome, we have the result that the market *will always be in equilibrium*. To see this, simply set $p_t^e = p_t^*$ in the market supply function $s(p_t^e)$ and solve for the resulting equilibrium price from

$$D(p_t^*, t) = s(p_t^*) \quad [\text{B.21}]$$

(where the time argument is included in the demand function to indicate that it may shift over time). This illustrates Muth's proposition: it is rational to take as one's expectation the predicted outcome of the market model. Here 'rational' means *profit maximizing*: no other farmer can increase profit by forming an expectation in any other way, given that all other farmers form their expectations this way (note the similarity to the concept of *Nash equilibrium* discussed in Chapter 12). This was certainly not true of the naive expectations assumption: if she believed that all other farmers simply extrapolated this period's price, the rational farmer would expand her potato acreage when current price is low and conversely.

The conclusion of the rational expectations hypothesis – that the market is always in equilibrium – implies that price fluctuations in the market are driven by shifts in the underlying supply and demand functions rather than by the expectations formation process as such. This is, at least in principle, a testable proposition. Certainly the rational expectations hypothesis is intellectually more appealing than the naive hypothesis, and we should note that forecasting and modelling markets are important areas of economic activity. Nevertheless, the hypothesis is a strong one and may not be appropriate for all markets.

Exercise 10B

1. Compare the stability properties of the Walrasian *TP* and Marshall's process when demand and supply curves both have positive slopes.
2. Do you think Cobweb cycles are most likely in the market for lettuce, coffee or lawyers?
3. *Correspondence principle: stability and comparative statics.* Consider a market with a continuous and differentiable supply function $s(p)$ and a continuous and differentiable demand function $D(p, a)$ where a is a shift parameter which increases demand: $D_a(p, a) > 0$. Assume that for each value of a there exists a p^0 and a p^1 such that $z(p, a) = D(p, a) - s(p) > 0$ for all $p \geq p^0$ and $z(p, a) < 0$ for all $p \leq p^1$. Thus there always exists an equilibrium price $p(a)$ satisfying $z(p, a) = 0$.

- (a) Show that an increase in a always increases the equilibrium price provided (i) the market adjustment process is the Walrasian TP and (ii) the initial equilibrium was stable. (Hint: differentiate the equilibrium condition totally with respect to a .)
- (b) What if the Marshallian adjustment process was used?
- (c) Show that under the Walrasian TP an increase in a always increases the equilibrium price even if the initial equilibrium was unstable. (Hint: draw some diagrams and use the intermediate value theorem. Differentiating the equilibrium condition will give a misleading result.)
- (d) Does the previous conclusion hold under the Marshallian adjustment process?
4. *Adaptive expectations.* Suppose that expectations are formed adaptively: $p_t^e - p_{t-1}^e = k(p_{t-1} - p_{t-1}^e)$ ($0 < k < 1$). Will the market converge to an equilibrium?

C. Long-run equilibrium

In Chapter 9 we saw that the *firm's* long-run supply curve is that part of its long-run marginal cost curve above its long-run average cost curve. There are several reasons why the market supply curve cannot be obtained simply by summing these supply curves:

- (a) A reason already familiar from the construction of the short-run market supply curve: as *all* firms vary output, we can expect input prices to change, thus causing each firm's cost curves to shift.
- (b) In addition to what are often called '*external pecuniary diseconomies*' described under (a), there may be *external technological diseconomies* – e.g. congestion, pollution – or *economies* – e.g. improvement of common facilities such as transport and communications – which also shift individual firms' cost curves as a result of expansion of scale by all firms.
- (c) *The number of firms in the market* will in general depend on the price – as price rises firms which previously found it unprofitable to produce the commodity now find it profitable, and so invest in capacity and add to output. A characteristic of a competitive market is that there are no barriers such as patents, legal restrictions, ownership of raw material sources, which impede the entry of new firms. A firm which at the going price just breaks even, with total revenue equal to long-run total cost (including the opportunity cost of capital and effort supplied by its owner(s)) is called a *marginal firm* at that price. One which makes an 'excess profit' (total revenue > total long-run opportunity costs) is called an *intra-marginal* firm, and one which would make a loss, but breaks even at a higher price, is called an *extra-marginal* firm. As price rises, marginal firms become intra-marginal and some extra-marginal firms enter.

Given this set of influences which determine how the rate of output changes with price as firms adjust capacity and entry or exit take place, it is by no means assured that the long-run market supply curve will be positively sloped (see Questions 1, 2). However, in

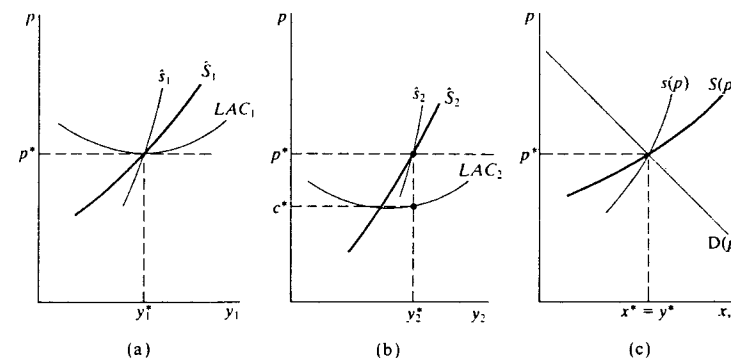


Fig. 10.7

Fig. 10.7(c) we assume this to be the case. $S(p)$ shows how the rate of output varies with price when capacity is adjusted and the number of sellers may change. It should be noted that underlying this curve is a possibly complex set of adjustments, and the transition from one point on the curve to another is not so smooth and effortless as the curve suggests. It should be interpreted as showing the aggregate output which will be forthcoming at each price *after* all these adjustments have been made. Or, alternatively, it shows the price at which a given number of firms would remain in the industry, maintain their capacity and supply in aggregate a given rate of output. The p -coordinate of any point y is then the *long-run* supply price of that rate of output.

The long-run equilibrium is shown in Fig. 10.7(c) as the point (y^*, p^*) . At this point firms are prepared to maintain the rate of supply y^* , and consumers are prepared to buy this output at price p^* . If, therefore, the short-run supply curve $s(p)$ was as shown in the figure, the short-run equilibrium we have earlier been examining would also be a long-run equilibrium. It would be maintained indefinitely in the absence of any change in demand, input prices or technology.

The other parts of the figure show the implications of the long-run equilibrium for two 'representative firms'. In (a) we show firm 1 as a marginal firm. At market price p^* it chooses a long-run profit maximizing scale of output y_1^* , and at that output, $p^* =$ its minimum long-run average cost. Firm 2, on the other hand, shown in (b) of the figure, is an intra-marginal firm; at its profit maximizing scale of output y_2^* , its long-run average cost $c^* < p^*$, and it makes an excess profit equal to $(p^* - c^*)y_2^*$. However, such 'excess profits', which may be earned temporarily, will not persist indefinitely, but rather should be regarded as true opportunity costs to the firm in the long run.

The argument goes as follows: the fact that the intra-marginal firm's average costs are lower than those of a marginal firm must reflect the possession of some particularly efficient input, for example especially fertile soil or exceptionally skilful management. Since these generate excess profits, we expect other firms to compete for them, so that after a period long enough for contracts to lapse, the firm which currently enjoys the services of these super-efficient inputs will have to pay them what they ask or lose them. The maximum these inputs can extract is the whole of the excess profit $(p^* - c^*)y_2^*$, and so what was a

profit during the period when the contract was in force becomes a true opportunity cost to the firm after that time. Such excess profits are therefore called *quasi-rents*, to emphasize that they are not true long-run excess profits, but merely rents accruing to the contractual property rights in certain efficient input services, which become transformed into costs in the long run. Once this transformation has taken place, the 'intra-marginal' firm's *LAC* curve will rise until its minimum point is equal to p^* . Hence in the long run all firms in the market will be marginal firms.

From Fig. 10.7 we can extract the three conditions which must hold in long-run equilibrium:

1. each firm in the market equates its long-run marginal cost to price, so that output maximizes profit;
2. for each firm price must equal long-run average cost (if necessary after quasi-rents have been transformed into opportunity costs) so that profits are zero and no entry or exit takes place;
3. demand must equal supply.

Conditions (1) and (2) then imply that each firm produces at the minimum point of its long-run average cost curve, as Fig. 10.7(a) illustrates. This is a strong result on the efficiency of the competitive market equilibrium, since it implies that total market output is being produced at the lowest possible cost.

The question arises of whether there is the possibility of discontinuity in this supply curve, of the kind we noted in the case of the short-run supply curve in Fig. 10.2(b), which could imply the non-existence of equilibrium. Such a case could arise as follows. Suppose that

- (a) all firms, whether currently in the market or not, have identical, *U*-shaped *LAC* curves as shown in Fig. 10.7(a);
- (b) input prices do not vary with industry output.

Then, there could be a discontinuity in the long-run market supply curve at price p^* in Fig. 10.7. At any price below p^* , all firms would leave the market, then market supply will fall to zero, while at price p^* planned market supply is y_1^* multiplied by the number of firms which are capable of producing the good with the given *LAC* curve. This discontinuity could be avoided if we assume there is some mechanism which selects potential suppliers in such a way as to ensure that any given market demand at price p^* is just met by the appropriate number of firms each producing at minimum long-run average cost. Then, the long-run market supply curve would be a horizontal line at price p^* : expansion of market output is brought about entirely by new entry rather than through output expansion by existing firms. Long-run equilibrium price cannot differ from p^* , and so is entirely *cost determined*. The level of demand determines only aggregate output and the equilibrium number of firms. Note that for a long-run market supply curve which is a continuous horizontal line we need the least-cost output of a firm (y_1^* in Fig. 10.7(a)) to be 'very small' relative to market demand, and the number of firms to be 'very large'.

More simply, if we assume that the technology of production is such that there is no range of outputs over which there are increasing returns to scale, then we can also eliminate any discontinuity in market supply. For example, if all firms experience decreasing returns to scale at all outputs then long-run average and marginal cost curves will be everywhere upward sloping and their horizontal sum (taking into account any input price effects), will have an intercept on the price axis.

Alternatively, if we assume all firms have identical production functions with *constant* returns to scale, and face identical (constant) input prices, then the long-run market supply curve is again a horizontal straight line. Each firm's long-run marginal cost curve is a horizontal line and coincides with its long-run average cost curve, and these are at the same level for all firms. Then, the only possible equilibrium price is given by this common marginal = average cost so that price is again completely cost-determined. Demand again determines only the aggregate equilibrium market output. Note that in such a market model, the equilibrium output of each firm, as well as the equilibrium number of firms producing in the market, are indeterminate.

Stability in the long-run

The analysis of the stability of long-run equilibrium in a competitive market must take into account the interaction between short- and long-run decisions of firms, the effects of new entry and the role of price expectations. We carry out the analysis for the case in which input prices increase with aggregate market output, and all firms have *U*-shaped cost curves. As shown in Fig. 10.8, the long-run market supply curve is upward sloping. It should be thought of as the locus of price-quantity points at which the long-run equilibrium conditions are satisfied: at each point, price = long-run marginal cost for each firm in the market, and no further entry or exit will take place at a given price because firms are just

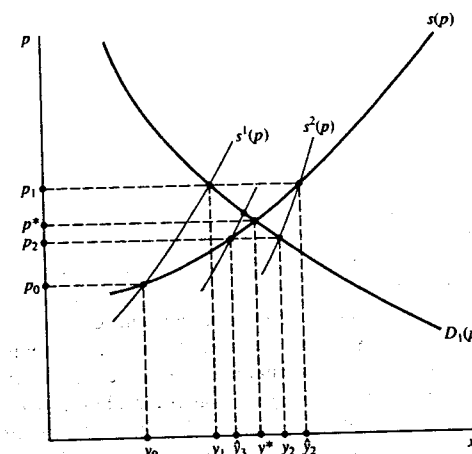


Fig. 10.8

breaking even at that price (given that the quasi-rents of intra-marginal firms have been transformed into opportunity costs). Thus corresponding to each point on the curve is a particular set of firms, each with a profit maximizing capacity and output level. As price rises, output along the curve increases as a result of both output expansion by existing firms and entry of new firms. However, the *actual time path of price and output may not coincide with the supply curve*. For that to happen, we again need the assumption of rational expectations, as we shall now see.

Suppose at year 0 the market is initially in long-run equilibrium at the price and output pair (p_0, y_0) in Fig. 10.8. However, in year 1 demand increases to $D_1(p)$. In the short-run – year 1 – output can only expand along the short-run supply function $s^1(p)$, determined by the short-run marginal cost functions of the firms already in the market (together with any effects of increasing input prices as analysed in section A). Thus price in year 1 is established as p_1 . Since p_0 corresponded to zero profits of the existing firms, p_1 must imply positive profits. The market is clearly not in long-run equilibrium. What happens next depends upon the assumption we make about price expectations formation.

Begin, as in the Cobweb Model of section B, with the assumption of naive expectations: all firms, whether currently in the market or contemplating entry, expect price p_1 to prevail next year, in year 2. The existing firms expand capacity and new firms enter and install capacity to the extent that planned market output expands to \hat{y}_2 , since this is the aggregate output corresponding to long-run profit maximization at price p_1 . But of course, when period 2 arrives, (p_1, \hat{y}_2) is not an equilibrium: price will have to fall to p_2 , where demand equals short-run supply as indicated by the short-run supply curve $s^2(p)$. This is determined by the short-run marginal cost curves of all firms in the market – initial incumbents and new entrants in year 2. If all firms again assume, naively, that p_2 will prevail in year 3, then capacity will be contracted and some firms will leave the market until \hat{y}_3 will be the aggregate market supply that will be *planned* for year 3! And so on. So, through time, under naive expectations, price will fluctuate around the equilibrium value p^* and, in the case illustrated in Fig. 10.8, eventually converges to it (in the absence of further demand change). The fact that capacity can only be adjusted ‘in the long run’ introduces the same kind of supply lag that we assumed for an agricultural market. The main difference is that here the short-run supply curve is positively sloped whereas in the Cobweb Model it was in effect vertical. The role of the long-run supply curve in the present analysis is to show how *future planned* output varies with the *expected future* price. Although the ultimate effect of the demand shift was to move the market from one point on the long-run supply curve to another, the actual time path of price and output through the adjustment process lies along the demand curve and describes a diminishing sequence of jumps from one side of the equilibrium point to the other.

However, our previous criticisms of the naive expectations assumption apply equally here. It is irrational for a profit maximizing firm to form its expectations in this way because then it is consistently sacrificing potential profits. Suppose instead that all firms have rational expectations, that is, they know the market model and use its prediction as their price expectation. Then, if the change in demand between periods 0 and 1 is unanticipated, the year 1 short-run equilibrium is at (p_1, y_1) as before, but now firms can predict the new long-run equilibrium price p^* . This is the only price with the property that the planned outputs which maximize profits at that price can actually be realized – i.e. sold – on the market next period. Hence existing firms will expand capacity and new

firms will enter so as to expand market output to y^* , and so the market moves to its full long-run equilibrium in year 2. If the change in demand had been fully anticipated at year 0, then the same argument leads to the conclusion that the market would move to its new long-run equilibrium in year 1. In that case, the market adjusts smoothly along its long-run supply curve to changes in demand.

Exercise 10C

1. Explain the shapes of the market supply curves in the cases:
 - (a) firms have identical constant returns to scale production functions and face identical constant input prices;
 - (b) firms have identical decreasing returns production functions and face identical constant input prices;
 - (c) as in (a) but with input prices increasing as market output increases.
2. Take an industry in which firms have identical *decreasing returns* production functions with input prices that fall as market output expands. Show that market equilibrium is fully determinate and a competitive market structure can be sustained, even if the market supply curve has a negative slope. Contrast this with the case in which firms have identical *increasing returns* production functions and input prices are constant.
3. Analyse the long-run adjustment process in markets of types (a), (b) and (c) of Question 1, first on the assumption of naive expectations, and second on the assumption of rational expectations.
4. Assume that all firms, incumbents and potential entrants, have identical U-shaped long run marginal and average cost curves. Analyse the process of adjustment from the initial equilibrium to a new equilibrium following an unexpected demand increase. (Assume constant input prices.) Explain why, under rational expectations, incumbent firms never change their long-run output and capacity.

Conclusions

The long-run market supply curve is a complex construction. Its slope and elasticity depend on: the nature of returns to scale in individual firms; the extent to which input prices vary with aggregate industry output; the existence of external technological economies and diseconomies of scale; and the flow of new entrants into the market. Moreover, the process of adjustment between points on the long-run supply curve, in response to demand shifts, may also be complex, and depends on the relation between short-run and long-run supply, on the one hand, and the nature of expectations formation, on the other. The analysis of long-run competitive equilibrium must then be treated with some care. On the other hand, the analysis of existence and stability of short-run equilibrium is relatively straightforward and has served to introduce some ideas which will be greatly extended in Chapter 16 on general competitive equilibrium.

References and further reading

For a discussion of the relationship between the Walrasian tâtonnement adjustment process and the Marshallian adjustment hypothesis see

D. G. Davis. 'A note on Marshallian vs. Walrasian stability conditions', *Canadian Journal of Economics and Political Science*, 29, 1963.

Stability analysis and its relationship with comparative statics is examined in

P. A. Samuelson. *Foundations of Economic Analysis*, Harvard University Press, Harvard, 1948, part II.

J. N. Bhagwati, R. A. Brechner and T. Hatta. 'The global correspondence principle', *American Economic Review*, 77, 1988, 124–32.

Rational expectations were introduced in

J. F. Muth. 'Rational expectations and the theory of price movements', *Econometrica*, 29, 1961, 315–35.

For a sceptical view see

R. Frydman. 'Towards an understanding of market processes: individual expectations, learning, and convergence to rational expectations equilibrium', *American Economic Review*, 72, 1982, 652–68.

A seminal article on disequilibrium price and quantity adjustment is

K. J. Arrow. 'Towards a theory of price adjustment', in M. Abramovitz et al., *The Allocation of Economic Resources*, Stanford University Press, Stanford, California, 1959.

CHAPTER 11**Monopoly****A. Introduction**

The assumption underlying the model of the competitive market, that buyers and sellers act as price-takers, is often not satisfied. Sellers perceive that the market price will vary with the amount of output they put on the market: buyers appreciate that an increase in their purchases will drive the price up. It is important to develop a set of theories about resource allocation under such conditions. This chapter and the next will be concerned with theories of the price-setting behaviour of *sellers*; parts of Chapter 14 will be concerned with analysis of cases in which *buyers* have influence over market price.

Given the basic assumption that a seller, usually taken to be a firm, perceives that market price and the quantity sold vary with each other, we obtain two separate types of theory, according to the assumption we make about the nature of the competitive relation with other sellers. If we assume that a firm recognizes that a change in its price will have a significant effect on the demand for the output of one or more other identifiable firms and that changes in these firms' prices have a similar effect on its own demand, then we have *oligopoly*. This will be examined in the next chapter. Here, we shall be concerned with the case in which no such *perceived interdependence* exists: there is no close competitive relationship between the firm in question and one or more other identifiable sellers. In setting its price the firm can ignore its effect on other sellers and we have *monopoly*. This characterization of monopoly must however be qualified: although at a given point in time there may be no close competitive relation with another firm, there may be a *potential* competitive relation, since in the long run other firms may be attracted to enter the market and sell in competition with the monopoly. This may place an important constraint on the behaviour of the monopolist. In the next two sections we proceed as if no such potential competition existed. In section D we examine the consequences of the threat of new entry.

Note that it is usual to define oligopoly and monopoly in terms of the size distribution of sellers in a market: monopoly is a 'single seller' of a good, and oligopoly is the case of a 'few sellers'. But consider the following cases:

- (a) A public utility may have a monopoly of the supply of electricity, yet in its sales to domestic consumers, there may be a close competitive relation with the firms which

sell oil, coal and gas. The relevant good here is 'energy' or 'heat', and the various 'monopolies' are in an oligopoly.

- (b) A cement manufacturer may be one of, say, five sellers in the nationwide cement industry. However, because of high transport costs, it may be able to vary its price over some range to buyers in a region around its cement works, without affecting the demand of any of the other sellers, and the same may be true of them. Each enjoys a local monopoly. Thus what is apparently an oligopolistic market is really a collection of monopoly sub-markets.
- (c) A restaurant is able to raise its prices relative to those of the other, say, forty restaurants in town, without losing all its customers to them; and is able to lower its prices relative to theirs, without taking all their customers. This is because of differences in quality, location, style of cooking, ambiance. If its gains or losses of customers are spread evenly over all other restaurants, then there is unlikely to be a perceived interdependence among the restaurants. Each restaurant could be regarded as a monopoly (although possibly with a very elastic demand curve). On the other hand, if the customer changes are concentrated on just one or two close rivals (the only other Chinese restaurants in town) then the restaurant is in an oligopolistic market.

The point of these examples is to show that the appropriate model to use depends not on the size distribution of firms in the market, but on the nature of the competitive relations between sellers. Indeed, the *appropriate definition of the 'market' depends on the nature of the competitive relations*, rather than the other way around. Our definitions of monopoly and oligopoly make this clear from the outset.

B. Price and output determination under monopoly

The theory of the firm (in the sense of Chapter 6) we use in the analysis of monopoly in this section is not essentially different from that which underlays the analysis of competitive markets. The firm is assumed to seek to maximize profit in a stable, known environment, with given technology and market conditions. We continue to assume diminishing marginal productivity and so, in the presence of fixed inputs, the firm's average and marginal costs will at some point begin to rise with the rate of output per unit time. However, we no longer assume that diminishing returns to scale set in at some point: we leave the question open, and permit any one of increasing, constant, or diminishing returns to scale to exist over the range of outputs we are concerned with. The essential difference with the competitive model is the assumption that the firm faces a downward sloping demand curve. We write its demand function in the inverse form:

$$p = D(q) \quad dp/dq < 0 \quad [\text{B.1}]$$

where p is price, q is output per unit time, and D is the demand function. We do not place restrictions on the second derivative of the function, but do require its first derivative to be negative.

The firm's total cost function is written as:

$$C = C(q) \quad C'(q) > 0 \quad [\text{B.2}]$$

where C is total cost per unit time. Marginal cost is always positive, but we do not place restrictions on the second derivative, the slope of the marginal cost curve. The profit function of the firm is:

$$\pi(q) = pq - C(q) \quad [\text{B.3}]$$

where π is profit per unit time. The output $q^* > 0$ maximizes the firm's profit only if it satisfies the conditions:

$$\pi'(q) = p + q dp/dq - C'(q) = 0 \quad [\text{B.4}]$$

$$\pi''(q) = 2 dp/dq + q d^2p/dq^2 - C''(q) < 0 \quad [\text{B.5}]$$

where [B.4] is the first-order and [B.5] the second-order condition. The term $(p + q dp/dq)$ is the derivative of total revenue pq with respect to q (taking account of [B.1]), and is *marginal revenue*. Thus, [B.4] expresses the condition of equality of marginal cost with marginal revenue. The term $(2 dp/dq + q d^2p/dq^2)$ is the derivative of marginal revenue with respect to output and so [B.5] is the condition that the slope of the marginal cost curve must exceed that of the marginal revenue curve at the optimal point. If marginal costs are increasing with output while, by assumption, marginal revenue is diminishing with output, [B.5] will necessarily be satisfied, since in that case:

$$C''(q) > 0 > 2 dp/dq + q d^2p/dq^2 \quad [\text{B.6}]$$

However, unlike the competitive case, the second-order condition may also be satisfied if $C''(q) < 0$ (see Question 2, Exercise 11B).

More insight into this solution can be gained if we write marginal revenue, MR , as:

$$MR = p(1 + (q/p) dp/dq) \quad [\text{B.7}]$$

Given the definition of the elasticity of demand from Chapter 3:

$$e = p dq/q dp < 0 \quad [\text{B.8}]$$

we can write as the relationship between demand elasticity and marginal revenue:

$$MR = p(1 + 1/e) \quad [\text{B.9}]$$

Clearly, $e < -1 \Rightarrow MR > 0$ while $e = -1 \Rightarrow MR = 0$, and $e > -1 \Rightarrow MR < 0$. Combining [B.9] with [B.4], we can write the condition for optimal output as:

$$p(1 + 1/e) = C'(q) \quad [\text{B.10}]$$

This equation then establishes immediately the two propositions:

- the monopolist's chosen price always exceeds marginal cost since its price elasticity is finite;
- optimal output is always at a point on the demand curve at which $e < -1$ (given that $C'(q) > 0$).

Because under competitive conditions market prices will be equal to each firm's marginal cost, the extent of the divergence of price from marginal cost under monopoly is often regarded as a measure of the degree of monopoly power enjoyed by the seller. From [B.10] we have:

$$\frac{p - C'(q)}{p} = \frac{-1}{e} \quad -\infty < e < -1 \quad [\text{B.11}]$$

where the left-hand side, the price marginal cost difference expressed as a proportion of the price, was defined by A. Lerner (1934) as an index of the degree of monopoly power. Thus, as $e \rightarrow -\infty$ (the competitive case) monopoly power tends to zero.

The equilibrium position of the firm implied by its choice of output q^* satisfying the above conditions, is illustrated in Fig. 11.1. In (a) of the figure the demand curve is $D(q)$ and the corresponding marginal revenue curve is MR . Given the marginal and average cost curves $C'(q)$ and AC , profit maximizing output is at q^* . Since this must be sold at a market-clearing price, choice of q^* requires the price $p^* = D(q^*)$. We could therefore regard the equilibrium position as a choice either of profit maximizing price p^* or of output q^* , since each implies the other. At output q^* , profit is the difference between total revenue p^*q^* and total cost $AC \cdot q^*$, and is shown by the area p^*abc in Fig. 11.1(a). In (b) of the figure the same equilibrium position is shown in terms of total revenue and cost curves. The total revenue curve is denoted pq , and its slope at any point measures marginal revenue at that output. Its concave shape reflects the assumption of diminishing marginal revenue. The total cost curve is denoted $C(q)$, and its convex shape reflects the assumption of increasing marginal cost. The total profit function is the vertical difference between these two curves, and is shown as the curve $\pi(q)$ in the figure. The maximum of this curve occurs at the output q^* , which is also the point at which the tangents to the total revenue and total costs curves respectively are parallel, i.e. marginal revenue is equal to marginal cost.

The 'supernormal profit', i.e. profit in excess of all opportunity costs (including a market-determined rate of return on capital which enters into determination of the average and marginal cost curves), given by the area $q^*(p^* - c)$, can be imputed as a rent to

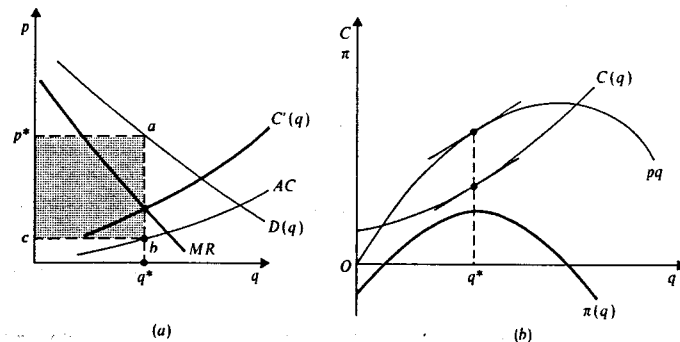


Fig. 11.1

whatever property right confers the monopoly power and prevents the new entry which would compete the profits away. It may be that this right is owned by an individual who leases it to the firm. If the supplier is rational and well-informed, she will bid up the price of the lease so as just to absorb the super-normal profit, and so the rent is transformed into an opportunity cost for the monopolist. This would be true, for example, if the monopolist rented a particularly favourable location. If the monopolist owns the property rights, then he can impute the profits as the return on this property right. Note that the identity of the owner of the right does not affect the price and output which will be set by the monopolist (since this is determined by the desire to maximize profit) but simply determines the division of the spoils. Note also that the term 'property right' is used here in its widest possible sense: it is meant to include not only the ownership of land, but also of brand names, public reputations, mineral rights, franchises, patents, in fact anything which allows its possessor to create and perpetuate monopoly power.

Exercise 11B

- The analysis in the text assumed implicitly that there were constraints on the firm's inputs, so that the short-run cost function of Chapter 8 was the relevant cost function for the firm's decisions. Extend the analysis to take account of the interaction of long-run and short-run decisions (cf. Chapter 9). Show that the firm will set output so that $SMC = MR$ in each period and will plan to produce next period where $SMC = LMC = MR$. (Assume no threat of entry exists.)
- Using diagrams analogous to those in Figs 11.1(a) and 11.1(b), illustrate cases in which there are increasing returns to scale.
- Show how a monopolist's price and output will be affected by:
 - an increase in demand;
 - a specific tax per unit of output;
 - a proportionate 'excess profit' tax.

Compare these comparative statics results with those for the competitive firm in Chapter 9.

- What method could you use to induce the monopolist to produce at the output at which $C'(q)$ cuts $D(q)$ (i.e. at which price = marginal cost) in Fig. 11.1(a)? Describe four methods, and assess their advantages and disadvantages. *NOTION 4*
- (a) Explain why it is meaningless to talk of 'the supply curve' of a monopolist. *yes*
(b) Is it also meaningless to talk of the demand curve of the monopoly for inputs it uses in production? *yes*
- Suppose that the right to be a monopolist is auctioned off by the government (for example the right to be the only petrol station on a stretch of motorway), what will be the monopoly price, output and profit if the right is given to the firm (a) making the highest money bid; (b) promising to sell petrol at the lowest price; (c) promising to pay the largest share of its revenue over to the government.

7. A monopoly produces two outputs which are interdependent in demand. Set up, solve and interpret a model of its profit-maximizing output choices.
8. A monopoly faces the inverse demand function $p = q^{-2}$ and has the total cost function $C = cq$. Analyse the problem which may arise in modelling its profit maximizing output choice. Under what condition would no such difficulty arise?
9. Explain why a monopolist with zero marginal cost would produce at an output at which $e = -1$. What is the value of the Lerner index at such a point?

C. Price discrimination

Price discrimination is the practice whereby different buyers are charged different prices for the same good. It is a practice which could not prevail in a competitive market because of *arbitrage*: those offered lower prices would resell to those offered higher prices and so a seller would not gain from discrimination. Its existence therefore suggests imperfections of competition, and we are particularly interested in its practice by a monopolist.

Third-degree price discrimination: market segmentation

Suppose that the monopolist can divide the market for his output into two sub-groups, between which arbitrage can be prevented at zero cost. To concentrate on essentials assume that the costs of supplying the two sub-markets are identical, so that any price difference between the sub-markets will arise from discrimination, not differences in, say, transport or distribution costs.

The monopolist knows the demand, and therefore marginal revenue, curves, for each group. Let q_1 and q_2 be the quantities sold to the first and second groups respectively, so that total output $q = q_1 + q_2$. Take some *fixed* total output level, q_0 , and consider the division of this between the two sub-markets in such a way as to maximize profit. Since the total production cost of q_0 is given, profit from the division of this between the two markets is maximized if revenue is maximized. But revenue is maximized only if q_1 and q_2 are chosen such that the marginal revenues in each sub-market are equal. To see this, let MR_1 be the marginal revenue in sub-market 1, and MR_2 that in 2. Suppose $MR_1 > MR_2$. Then it would be possible to take one unit of output from market 2, and put it in market 1, with a net gain in revenue of $MR_1 - MR_2 > 0$. As long as the marginal revenues were unequal such possibilities for increasing revenue, and therefore profit, would exist. Hence a necessary condition for a profit maximizing allocation of any given total output between the two markets is that marginal revenues in the markets be equal.

In determining the optimal total output level, we are on familiar ground. If $MR_1 (= MR_2)$ differed from marginal cost, it would be possible to vary output in such a way as to increase total profit: by increasing output when $MR_1 > MC$, and reducing it in the converse case. Hence a necessary condition for maximum profit is that $MC = MR_1 = MR_2$.

Now let e_1 and e_2 be the price elasticities of demand in the respective sub-markets. Then, the basic relation given in [B.9] applies in this case, so that:

$$MC = p_1(1 + 1/e_1) = p_2(1 + 1/e_2) \quad [C.1]$$

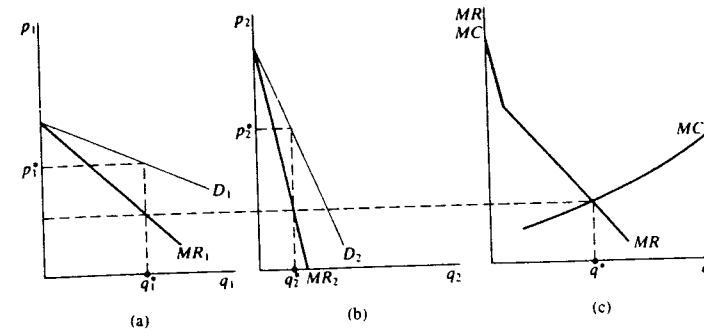


Fig. 11.2

From the second equality in [C.1] we have:

$$\frac{p_1}{p_2} = \frac{1 + \frac{1}{e_2}}{1 + \frac{1}{e_1}} \quad [C.2]$$

If $e_1 = e_2$, then clearly $p_1/p_2 = 1$, and there is no discrimination, but there will be as long as the elasticities are unequal at the profit maximizing point. Moreover, if $e_1 < e_2$, then from [C.2] $p_1 < p_2$, and conversely. (Remember $e_i < 0$.) We conclude that in maximizing profit the monopolist will always set a higher price in the market with the lower elasticity of demand.

The analysis is illustrated in Fig. 11.2. In (a) of the figure are the demand and marginal revenue curves for sub-market 1 and in (b) those for 2. The curve MR in (c) is the *horizontal sum* of the MR_1 and MR_2 curves, and therefore has the property that at any total output, q^0 , the two output levels q_1^0 and q_2^0 which have the same marginal revenues in the sub-markets as that at q^0 , sum exactly to q^0 , i.e. $q_1^0 + q_2^0 = q^0$. The horizontal summation therefore reflects the first condition derived above, that any total output must be divided between the sub-markets in such a way as to equalize their marginal revenues. The profit maximizing level of total output is shown at q^* , where $MC = MR$. To divide q^* optimally between the sub-markets, we simply take q_1^* and q_2^* , the sub-market outputs which have the appropriate marginal revenues, and which by construction must sum to q^* . We see immediately from the figure that since demand for q_2 is less elastic than that for q_1 , we have $p_2^* > p_1^*$.

First-degree discrimination

The above analysis dealt with third-degree price discrimination where the monopolist had some information on the basis of which he could partition buyers into sub-markets and prevent arbitrage between the sub-markets. This, as the name suggests, is in contrast to

two other types of price discrimination:

1. *first-degree price discrimination*, where the monopolist is able to identify the demand of each individual buyer and prevent arbitrage among all buyers;
2. *second-degree price discrimination*, where the monopolist knows the demand characteristics of buyers in general, but does not know which buyer has which characteristics.

In first-degree price discrimination the monopolist can extract *all* the consumer surplus of each buyer. An interesting aspect of this case is that total output of the good is at the level at which each buyer pays a price equal to marginal cost. Thus we have the 'competitive outcome': monopoly does not distort the allocation of resources and so, in the terminology of Chapter 17, we have a Pareto efficient outcome, with the monopolist expropriating *all* the gains from trade. Any objection to monopoly in this case therefore would have to be on grounds of equity – fairness of the income distribution – rather than efficiency.

In the second case, the obstacle to price discrimination is that if one type of buyer is offered a more favourable deal than another type, and the monopoly is not able to identify a buyer's type, then all buyers will take the most favourable deal. The solution for the monopolist is to offer alternative deals which satisfy a self-selection constraint: a given deal will be preferred to all others by, and only by, the type for which it is designed.

In the rest of this section we explore both these forms of price discrimination with a simple model. We assume:

- (a) two types of buyer in the market, with n_1 buyers of the first type and n_2 buyers of the second;
- (b) a buyer's type is determined by her utility function, which for each type of buyer takes the *quasi-linear* form

$$u_i = U_i(x_i) + y_i \quad i = 1, 2 \quad [\text{C.3}]$$

where x_i is the monopolized good, and y_i is a composite commodity representing all other goods;

- (c) type 2 buyers have a stronger preference for the good in the sense that for any x

$$U_2'(x) > U_1'(x) > 0 \quad [\text{C.4}]$$

- (d) the buyers have identical incomes M , and so if $x_1 = x_2 = 0$, then $y_1 = y_2 = M$ (the price of the composite commodity is the same for all consumers and is set at unity);
- (e) $U_i(0) = 0$ and $U_i''(x) < 0$: buyers have diminishing marginal utility;
- (f) the monopolist produces at a constant marginal (= average) cost c .

Recall from question 2, Exercise 4c that a quasi-linear utility function implies that a consumer's indifference curves in the x, y plane are vertically parallel, and there is a zero income effect for good x . The consumer's choice problem is:

$$\max_{x_i, y_i} U_i(x_i) + y_i \quad \text{s.t. } px_i + y_i = M - F \quad [\text{C.5}]$$

First-order conditions include

$$U_i' - \lambda p = 0 \quad [\text{C.6}]$$

$$1 - \lambda = 0 \quad [\text{C.7}]$$

Hence $U_i'(x) = \lambda p = p$, yielding demand functions

$$x_i = U_i'^{-1}(p) = x_i(p) \quad [\text{C.8}]$$

$$y_i = M - F - px_i(p) \quad [\text{C.9}]$$

The indirect utility function is

$$v_i(p, F) = U_i(x_i(p)) + M - F - px_i(p) \quad [\text{C.10}]$$

Here, p is the price the monopolist charges, and $F \geq 0$ is a *fixed charge* that the monopolist may set for the right to buy the good at price p (examples of such fixed charges are telephone rentals, entrance charges to amusement parks, subscription fees to a book or wine club). Of particular interest are the derivatives:

$$\frac{\partial v_i}{\partial p} = U_i' x_i' - (x_i + p x_i') = -x_i; \quad \frac{\partial v_i}{\partial F} = -1 \quad [\text{C.11}]$$

where the result for $\partial v_i / \partial p$ is simply Roy's identity. In Fig. 11.3a, we show the 'reservation indifference curves' \bar{u}_i for each of the two types of consumers. Since they have the same income M , they are at the same point when consuming no x , but assumption (c) implies that a type 2 indifference curve is steeper than that of a type 1 at every x (since $MRS_{xy}^i = -dy_i/dx_i = U_i'(x)$). The budget line marked c in the figure corresponds to $p = c$, so that x_i^c are the respective consumers' demands at that price. In (b) of the figure we show the demand curves derived from these reservation indifference curves. Because of the quasi-linearity assumption, these are both Hicksian and Marshallian demand curves, and the area under each between prices p_i^0 and $p = c$ gives the type's compensating variation, or maximum willingness to pay for the right to buy x at price c . These consumer surpluses are denoted by S_i , and correspond to the distances on the y axis shown in (a) of the figure.

We now show that under first-degree price discrimination the monopolist's optimal policy is to set a price for each type equal to c , and to set a fixed charge $F_i = S_i$, $i = 1, 2$. That is, the monopolist sells at marginal cost and sets separate fixed charges equal to the total willingness of each type to pay. This requires first that he knows the type of each buyer, and so can prevent a type 2 buyer taking advantage of the lower type 1 fixed charge. Second, he must be able to prevent arbitrage and stop a type 1 buyer reselling to a type 2 buyer at some price between c and $F_2/x_2^c + c$, which is the average price per unit a type 2 buyer pays in this solution.

The idea underlying this policy can be seen in Fig. 11.3b. If the monopolist sets $p = c$ to both types and extracts the total surplus his profit is $S_1 + S_2$. If on the other hand he sets a higher price, say $p' > c$, although he makes a profit on each unit he sells, the sum of these profits and the *remaining* consumer surpluses is less than $S_1 + S_2$ by the sum of the two shaded triangles. It pays him to expand output and lower price as long as $p > c$ because his own profit increases precisely by the difference $p - c$, which he can recover through the fixed charge. He will not set a price such as $p'' < c$, because the extra surplus

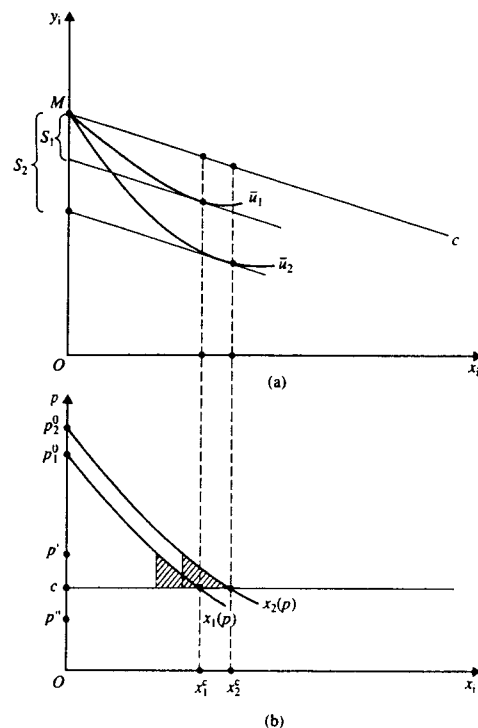


Fig. 11.3

he can recover falls short of the extra cost he incurs. And clearly it would never be worthwhile to set a fixed charge $F_i > S_i$ for any p , because then he sells nothing to type i .

We can derive this result more formally as follows. The monopolist's total profit is

$$\pi = n_1[p_1 x_1(p_1) + F_1] + n_2[p_2 x_2(p_2) + F_2] - c[n_1 x_1(p_1) + n_2 x_2(p_2)] \quad [\text{C.12}]$$

In maximizing this, he must take account of the fact that he cannot be too greedy: he must not offer a deal which is worse for each consumer than not buying the good at all. We can express this by the *reservation constraints*

$$v_i(p_i, F_i) \geq \bar{u}_i \quad i = 1, 2 \quad [\text{C.13}]$$

where, recall, \bar{u}_i is the utility i obtains by buying none of good x . With β_i as the Lagrange

multiplier on these constraints, optimal p_i and F_i are defined by

$$n_i(x_i + p_i x'_i - c x'_i) + \beta_i \partial v_i / \partial p_i = 0 \quad i = 1, 2 \quad [\text{C.14}]$$

$$n_i + \beta_i \partial v_i / \partial F_i = 0 \quad i = 1, 2 \quad [\text{C.15}]$$

$$v_i(p_i, F_i) \geq \bar{u}_i \quad \beta_i \geq 0 \quad \beta_i[v_i - \bar{u}_i] = 0 \quad i = 1, 2 \quad [\text{C.16}]$$

From [C.15] we see that non-zero n_i and $\partial v_i / \partial F_i$ imply $\beta_i > 0$ and so [C.16] implies $v_i = \bar{u}_i$. Both types of consumers receive only their reservation utilities. Then, using [C.11] and [C.15] we have $\beta_i = n_i$ and

$$n_i(x_i + p_i x'_i - c x'_i) - n_i x_i = 0 \quad [\text{C.17}]$$

implying

$$p_i = c \quad [\text{C.18}]$$

The value of F_i then satisfies $v_i(c, F_i) = \bar{u}_i$ and so must be equal to consumer surplus S_i at price c .

We could interpret third-degree price discrimination (analysed in the first part of this section) as the case in which the monopolist can identify each buyer's type and prevent arbitrage between types, but for some reason cannot set fixed charges. He must set a constant price per unit to all buyers of a given type. (See Question 9, Exercise 11C.) Then, profit maximization implies a price to each type which is above marginal cost, as we saw earlier. Clearly, the monopolist's profits are lower than under first-degree price discrimination. Interestingly, buyers are better off under third-degree price discrimination since, although they face a higher price and so consume less, they retain some consumer surplus and are on an indifference curve that must be higher than their reservation indifference curve (use Fig. 11.3).

Second-degree price discrimination

In the case of second-degree price discrimination, the monopolist is assumed to be unable to determine the type of the buyer before she has purchased the good. In that case if he offered any buyer the option of either (c, S_1) or (c, S_2) , every type 2 buyer (as well as every type 1 buyer) would choose (c, S_1) . Can the monopolist do better than this by offering options chosen so that only a buyer of type i would want to choose the option designed for her type? In other words, can the monopolist do better by inducing buyers to reveal their type by 'self-selecting' the appropriate deal?

Assume that the monopolist knows the number of buyers of each type, n_i , and can specify in a contract *both* the quantity of output he will supply to a buyer *and* the total charge for that output. That is, a contract is a pair (x_i, F_i) . This implies a price per unit $p_i = F_i/x_i$ and the contract may even be formally expressed as some combination of a fixed charge and constant price per unit, as in a two-part tariff. The important point is that the consumer is offered a *quantity* and a fixed charge, and not a *price* and a fixed charge. We shall see the reason for this at the end of the following analysis.

The monopolist's profit is

$$\pi = \sum_{i=1}^2 n_i(F_i - cx_i)$$

We again have the reservation constraints, since buyers always have the option of refusing a contract. These are now written in terms of direct utilities, to reflect the fact that quantities are being specified:

$$U_i(x_i) + M - F_i \geq \bar{u}_i \quad i = 1, 2$$

where we use the fact that $y_i = M - F_i$. There are also *self-selection* constraints which ensure that each type chooses the appropriate deal. We write these as

$$U_1(x_1) - F_1 \geq U_1(x_2) - F_2$$

$$U_2(x_2) - F_2 \geq U_2(x_1) - F_1$$

(M cancels out in these expressions).

If (x_i, F_i) satisfies these constraints, it will only be accepted by type i . (We assume, to be able to have a *closed* feasible set, that if a buyer is indifferent between the two deals she takes the one appropriate to her type.)

In principle we now solve for x_i, F_i by maximizing π subject to [C.20]–[C.22]. However, the first-order conditions for this would not be instructive. Instead, we first show that, in any optimal solution, (a) the *reservation* constraint for a type 2 buyer, and (b) the *self-selection* constraint for a type 1 buyer are non-binding. They can be dropped from the problem thus simplifying the derivation of the optimal contract.

We show this in Fig. 11.4, which reproduces the reservation indifference curves from Fig. 11.3a.

(a) Type 2 buyers must be offered (x_2, F_2) such that $u_2 > \bar{u}_2$. To see that, note that type 1 buyers must be offered a contract (x_1, F_1) that puts them on or above \bar{u}_1 . But since \bar{u}_1

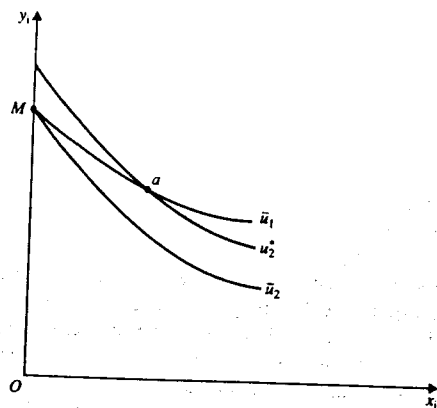


Fig. 11.4

above \bar{u}_2 , such a deal must always be better for type 2 buyers than any contract (x_2, F_2) that puts them on \bar{u}_2 . So only a point above \bar{u}_2 can satisfy their self-selection constraint.

(b) Type 1 buyers will always strictly prefer their deal to that offered to type 2 buyers, if an optimal solution. Suppose the optimal deal offered to type 1 buyers is at a point a (it is not relevant to the present argument that a is on \bar{u}_1 , but we show below that this must be so). Then the deal offered to type 2 buyers must lie on the type 2 indifference curve passing through a , labelled u_2^* . If it were below this, type 2 buyers would prefer a ; if above, the monopolist is being needlessly generous to type 2 buyers because, at any given x_2 , he could increase F_2 (move vertically downward in the figure) without violating either the reservation or self-selection constraints. (This incidentally established that the self-selection constraint for type 2 buyers is strictly binding, as we verify later.) Now if the deal offered to type 2 buyers were on u_2^* at a point to the left of a , this would be preferred by type 1 buyers and this contradicts the optimality of a . It is easy to show that point a itself could not be offered to both types of buyer in equilibrium (see Question 8, Exercise 11C). This leaves only points on u_2^* to the right of a as possible deals to be offered to type 2 buyers, and since these must be strictly below \bar{u}_1 , the type 1 self-selection constraint is non-binding. This argument also establishes that at an optimum $x_2 > x_1$.

As a result of these arguments, the monopolist's problem is to find $(x_1, F_1), (x_2, F_2)$ to maximize π in [C.19] subject only to [C.20] with $i = 1$, and [C.22]. Using β_1 and μ_2 for the Lagrange multipliers on [C.20] and [C.22], the first-order conditions are

$$-n_1c + \beta_1 U'_1(x_1^*) - \mu_2 U'_2(x_1^*) = 0 \quad [\text{C.23}]$$

$$-n_2c + \mu_2 U'_2(x_2^*) = 0 \quad [\text{C.24}]$$

$$n_1 - \beta_1 + \mu_2 = 0 \quad [\text{C.25}]$$

$$n_2 - \mu_2 = 0 \quad [\text{C.26}]$$

$$U_1(x_1^*) + M - F_1^* - \bar{u}_1 \geq 0 \quad \beta_1 \geq 0 \quad \beta_1 [U_1 + M - F_1^* - \bar{u}_1] = 0 \quad [\text{C.27}]$$

$$U_2(x_2^*) - F_2^* - U_2(x_1^*) + F_1^* \geq 0 \quad \mu_2 \geq 0 \quad \mu_2 [U_2 - F_2^* - U_2 + F_1^*] = 0 \quad [\text{C.28}]$$

From [C.26] and [C.28] we see that the type 2 self-selection constraint must bind, and from [C.25] and [C.27] that the type 1 reservation constraint must bind. Substituting for μ_2 in [C.24] gives

$$U'_2(x_2^*) = c \quad [\text{C.29}]$$

implying $x_2^* = x_2^c$, so that type 2 consumption is exactly that under first-degree price discrimination. Then, substituting for β_1 and μ_2 in [C.23] gives

$$U'_1(x_1^*) = \frac{n_1c}{n_1 + n_2} + \frac{n_2}{n_1 + n_2} U'_2(x_1^*) \quad [\text{C.30}]$$

Recall that we established in Fig. 11.4 that we must have $x_2^* > x_1^*$, so that $U'_2(x_1^*) > U'_2(x_2^*) = c$, given diminishing marginal utility. Thus, writing $U'_2(x_1^*) \equiv c + \delta$,

where $\delta > 0$, we have

$$U'_1(x_1^*) = c + \frac{n_2 \delta}{n_1 + n_2} \quad [\text{C.31}]$$

implying that $x_1^* < x_1^c$, so that type 1 buyers consume less than under first-degree price discrimination. The optimal values F_1^* and F_2^* then follow from solving the constraints as equalities with the optimal x_1^* inserted. We know that F_1^* will leave type 1 buyers with their reservation utilities, while F_2^* is such that type 2 buyers retain some consumer surplus. It follows that compared to first-degree price discrimination, type 1 buyers are neither better nor worse off, type 2 buyers are better off, and the monopoly makes less profit.

The optimal second-degree price discrimination equilibrium is illustrated in Fig. 11.5. The contracts are (x_1^*, F_1^*) and (x_1^c, F_2^*) . The two most interesting aspects of the solution are first, that $x_1^* < x_1^c$, and second that $x_2^* = x_2^c$. These can be rationalized as follows. At any x_1 , the total net surplus can be expropriated from type 1 buyers since they can be held to their reservation constraint. Suppose x_1 were set at x_1^c . This cannot be optimal. The contract for type 2 buyers would have to be a point on the indifference curve \bar{u}_2 , as shown in Fig. 11.5. Now consider a small reduction in x_1 from x_1^c . Since at x_1^c net surplus is at a maximum, this results in a change in net surplus from type 1 buyers of just about zero. On the other hand, it permits a downward shift in the indifference curve on which type 2 buyers can be placed, and, at any x_2 this results in a strictly positive gain in net surplus to the monopolist. Thus it pays to reduce x_1 below x_1^c . Of course, for further reductions in x_1 the monopolist will lose some net surplus from type 1 buyers, but this must be traded off against the gain in surplus from type 2 buyers, and the optimum, x_1^* , just balances these at the margin.

To see why $x_2^* = x_2^c$, note that it pays the monopolist to maximize the net surplus of type 2 buyers with respect to output, since this then maximizes the value of F_2 that can be set, subject to the constraint that type 2 buyers would not prefer the type 1 contract.

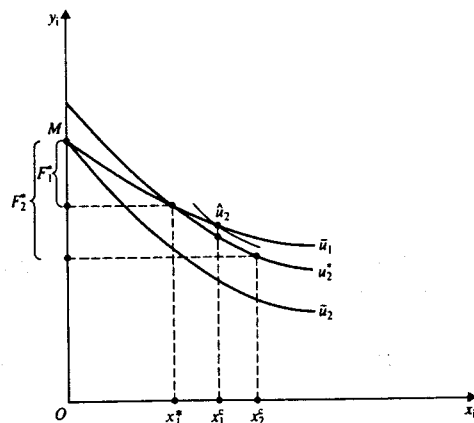


Fig. 11.5

There is a qualification to the condition in [C.31]. Note that as n_1 falls, given n_2 , x_1^* must also fall. It is then possible, for suitably small n_1 , that [C.31] cannot be satisfied for any $x_1 > 0$, in which case F_1 is set sufficiently high that no type 1 buyers enter the market. The monopolist then knows that the only buyers in the market are of type 2, and so he can extract all their consumer surplus, with $F_2^* = S_2$. In terms of Fig. 11.5, u_2^* becomes \bar{u}_2 . The intuitive explanation is that when the proportion of type 1 buyers is sufficiently small, the loss in total profit from reducing x_1 , and the corresponding extracted surplus, is small relative to the gain from being able to extract more surplus from type 2 buyers. The equilibrium position in Fig. 11.5 depends on the proportions of buyers of the two types as well as on the shapes of the indifference curves and the value of c .

The importance of the specification of quantities in the contract can be seen if we consider the *two-part tariffs* implied by the equilibrium in Fig. 11.5. If type 1 buyers took a contract in which they paid a fixed charge C_1^* and then a price per unit of $p_1^* = U'_1(x_1^*)$, then they would choose consumption x_1^* and pay precisely $C_1^* + p_1^* x_1^* = F_1^*$. Likewise, if type 2 buyers were set a fixed charge C_2^* and paid a price per unit $p_2^* = U'_2(x_2^*) = c$ then they would choose to consume x_2^* and pay in total $F_2^* = C_2^* + c x_2^*$. If the monopolist made these contracts available to all buyers and did not restrict the quantity that could be bought, Fig. 11.6 shows that the self-selection constraint would be violated. Type 2 buyers would clearly choose a type 1 contract, which would dominate the contract (x_2^*, F_2^*) , although type 1 buyers still prefer their own contract. On the other hand, if the monopolist specified contracts of the form: a fixed charge C_1^* and a price per unit p^* , up to a maximum of x_1^* , units of consumption, or a fixed charge C_2^* and a price of c for any amount of consumption, then the self-selection constraints would continue to hold. In fact, the tariffs or price schedules that firms with market power offer often do specify maximum consumption quantities as well as fixed and variable charges.

A note on terminology. *Linear pricing* refers to the case in which a buyer is charged a fixed price p per unit bought, so that her total expenditure is $E = px$, a linear function. A

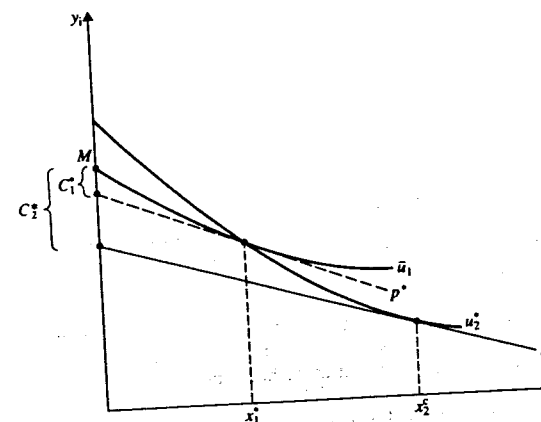


Fig. 11.6

two-part tariff consists of a fixed charge C and a fixed price p per unit bought, so that total expenditure is the affine function $E = C + px$. In this case, the average price per unit, $p + C/x$, is a non-linear, decreasing function of the quantity bought. In Fig. 5.11, the implied unit price F_1^*/x_1^* to each type of buyer will not be the same, implying a kind of non-linearity in the way in which unit price varies with quantity bought. Thus this kind of price discrimination, as well as two-part tariffs, falls under the general heading of 'non-linear pricing'.

To summarize this discussion of price discrimination: if a seller can identify each buyer's type (her demand function), and prevent arbitrage between types, then he maximizes profit by offering a two-part tariff consisting of a unit price equal to marginal cost c , and a fixed charge which expropriates all the consumer surplus of the given type. If a seller cannot identify a buyer's type, he must offer optional contracts: a high demand type will choose a contract which offers a unit price equal to marginal cost and a fixed charge which leaves her with some consumer surplus; a low demand type will choose a contract which offers a higher price up to a quantity maximum (x_1^*) and a lower fixed charge which nevertheless appropriates all her consumer surplus. Alternatively, the contracts may simply specify a quantity supplied and a total charge for that quantity. The aim is to prevent high demand buyers pretending to be low demand buyers, and taking the contract the latter would be offered under first-degree price discrimination, by making the low demand buyers' contract less attractive to the high demand buyers. Finally, if a buyer's type can be identified and arbitrage between types can be prevented, but the seller is constrained to use linear pricing, we have third-degree price discrimination. It is left as an exercise to show that the monopolist's profit falls as we move from first- through second- to third-degree price discrimination.

Exercise 11C

- 1.* Academic journals charge different subscription rates to institutions (college libraries, etc.); individual academics; and students. Explain this in terms of the theory of price discrimination. What would you predict about the pattern of relative subscription rates across these groups? Some journals are owned by profit maximizing firms and others by learned societies. What difference, if any, would you expect this to make to (a) the level of their rates and (b) the pattern of price discrimination?
2. Why are sparking plugs sold to car manufacturers as 'initial equipment', to be installed in new cars, at a price just about equal to average production cost, and sold to retailers and garages, for replacement purposes, at a price several times greater than average production cost?
3. Why are the fees charged by solicitors and estate agents, for services provided in buying and selling houses, expressed as percentages of the house price, even though the cost of the services involved is independent of the house price?
4. Why do firms sometimes offer quantity discounts ('one packet for 50p, two for 90p')?

- 5.* A firm which monopolizes one good may sometimes insist that people wishing to buy that good must also buy their requirements of some other good, which would otherwise be competitively produced, from the monopolist. (Examples have included Kodak and IBM.) Why is this *full line force* profitable given that the monopolist can charge a monopoly price for the monopolized good?
6. *Multinational firm.* A monopolist sells its output in Japan and in America. It also has a factory in both countries. Its profit maximization problem is to choose the amounts produced and the amounts sold in each country.
 - (a) Solve its problem diagrammatically.
 - (b) Suppose that the dollar is devalued against the yen. What effect will this have on the firm's decisions if it is (i) Japanese owned, (ii) American owned?
7. In the model of second-degree discrimination impose the constraints that the fixed charges must be zero, hence obtaining the case of third-degree discrimination. Use the resulting first-order conditions to confirm the diagrammatic analysis of third-degree discrimination.
- 8.* Show that the monopolist's profit falls as he moves from first- to second- to third-degree price discrimination.
9. A monopolist has two sub-market demand functions $p_i = a_i - b_i q_i$ and the total cost function $C = c(q_1 + q_2)$ where $c > 0$ is a constant. Compare prices, outputs and profits for the cases in which he does and does not practice price discrimination. Give an expression for the maximum cost the monopolist would incur to be able to prevent arbitrage.
10. Show that, on the assumptions of this section, under second-degree price discrimination different types would be offered different contracts.
11. *Self-selection by quality difference.* Monopolists often produce high and low quality goods and set prices such that the price differential between the high and low quality exceeds the additional cost of the higher quality version. Examples include first and tourist class seats on airlines, hardcover and paperback books. Adopt the analysis of second-degree price discrimination to explain this practice.

D. Entry

The analysis in the previous two sections has suggested that in the absence of competition, or the threat of it, from other sellers, a monopolist will earn excess profits. We cannot ignore the possibility of new entry, however. The existence of excess profits will be an attraction for other sellers. Monopoly power might therefore sow the seeds of its own destruction, and we expect a rational monopolist to take this into account. We now analyse some implications of the possibility of new entry for the behaviour of the monopolist.

The first question is, do *barriers to entry* exist? We distinguish between an *absolute entry barrier*, which rules out, over some time horizon, all new entry whatsoever; and a *relative entry barrier*, which places a new entrant at a disadvantage, but not an insurmountable

one. An absolute entry barrier may arise out of some legal impediment, such as a patent or statutory monopoly right, or out of the exclusive ownership of some resource which is indispensable for production. In that case, we interpret the monopoly profits, which will continue for as long as the absolute barrier exists, as rents accruing to the monopolist's holding of the legal rights or privileges.

Relative entry barriers may arise out of: capital market imperfections; specific cost advantages; and consumer loyalty. Capital market imperfections imply that different borrowers pay different interest rates, and also that the interest rate increases with the amount borrowed. This means that an entrant may have to pay a higher interest rate than the well-established monopolist, particularly if the lenders regard the entry as a risky proposition. It takes on particular force if there are significant economies of scale in production of the monopolized good. If the entrant sets up production on a scale smaller than that at which long-run average costs are at a minimum, then she will incur average costs which exceed those of the monopolist (assuming the latter is producing at minimum long-run average costs). On the other hand, if she enters on a scale large enough to achieve minimum average costs, the capital expenditure required may be very large, and may again involve her in higher interest costs than those incurred by the monopolist. Indeed, if there is capital rationing in the capital market, then the entrant may not be able to obtain the amount of funds she would require to set up on the optimal scale. If, on the other hand, the capital market were perfect, then both the monopolist and potential entrants would borrow at the same interest rate, and the entrants could borrow as much as they wished at the going rate, so the scale of capital expenditure required would be irrelevant.

'Specific cost advantage' is a cover-all term for things like superior location, availability of marketing outlets, advantageous input supplies, information, expertise and contacts, which are enjoyed by an established firm and make its costs lower, other things being equal, than those of a firm new to the market.

Consumer loyalty, built up and reinforced by advertising, and strengthened perhaps by innate conservatism and risk aversion of buyers, may impose on an entrant higher costs of advertising, packaging, sales promotion and product quality. In order to get her product known and accepted she will have to spend more on these than the established monopolist, at least in the initial stages.

Each of these relative entry barriers can be converted into a cost, and incorporated into the long-run cost curve of the entrant, which would then lie above that of the monopolist. A consequence of this would be that, since entry takes place only as long as the entrant anticipates excess profits, positive monopoly profits are not a sufficient condition for new entry. Then, any excess profits which remain to the monopolist could be imputed as rents to the factors which create the relative entry barrier.

Note that a relative entry barrier depends on the characteristics of both the monopoly and the potential entrant. In general, we might expect different potential entrants to have different long-run average costs of producing the good supplied by the monopolist. For example, a large firm, well-established in a market which is closely related to the monopolized one, may have little difficulty in raising cheap capital, may possess information about the market, and may be able to use its reputation in its existing markets to overcome consumer resistance in the new one. Indeed, a great deal of the 'new entry' which takes place is in the form of diversification and integration by already established firms. Thus it is not possible to gauge the extent of relative entry barriers by reference to

the characteristics of the monopoly alone; it is also necessary to take account of the characteristics of potential entrants.

Limit pricing

The monopolist could adopt a pricing policy which makes new entry unattractive: he may set a *limit price*. Clearly, an effective limit price would be one which was equal to the monopolist's own long-run average cost, since in that case no excess profit would be made and so no signals would go out to other sellers that opportunities exist for excess profit. However, we assume that the monopolist wishes to maximize profit, *subject to the constraint* that no other seller will find it profitable to enter the market. The question then is whether there exists a limit price which yields positive excess profits to the monopolist, and, if so, how is it determined? We now show that *provided the monopolist adopts or threatens to adopt the appropriate post-entry response* and that *this is believed by the potential entrant*, a limit price exists which yields positive excess profits. This is true even in the absence of relative entry barriers, while the existence of such barriers would increase the profitability of the limit-pricing strategy.

Assume:

- the monopolist knows the long-run average cost curve on which a potential entrant will operate (possibly, though not necessarily, because it is identical with his own);
- he also knows the market demand curve;
- the entrant's output would be undifferentiated from his own, so that both firms' outputs must sell at the same price;
- economies of scale exist over a significant range of the entrant's long-run average cost curve.

Fig. 11.7 sets out the analysis. D is the market demand curve for the monopolist's output, and AC_E is the potential entrant's long-run average cost curve (incorporating assumption (d)). The entry-excluding price which maximizes the monopolist's profit is p_m^*

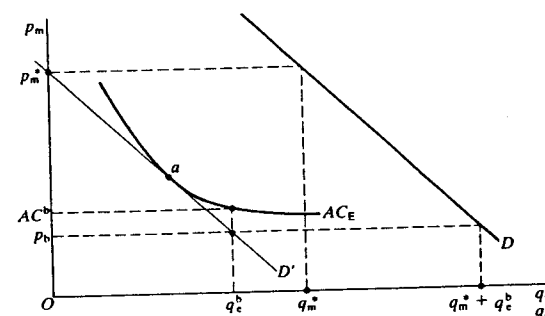


Fig. 11.7

implying an output of q_m^* . It is always assumed that the entrant's costs are such as to make the highest possible limit price less than the price which maximizes the monopolist's short-run profits, otherwise the problem is trivial. p_m^* is the optimal limit price.

The reasoning is as follows: consider the demand curve D' , which is found by shifting D leftward and parallel to itself until it is just tangent to the entrant's average cost curve. This tangency point is labelled a . The price p_m^* is found as the point at which D' cuts the vertical axis. Suppose that the entrant is made to believe that if she enters the market, the existing seller will maintain his output at q_m^* , so that total market output will be q_m^* plus the entrant's output. Then, she will perceive that price must fall along the portion bD of the market demand curve, to an extent dependent on her own output. In other words, the line p_m^*D' is effectively the entrant's perceived demand curve, since it shows the price at which she can sell the various amounts of outputs she may want to put on to the market. But since this demand curve lies nowhere above the entrant's long-run cost curve, she will see that she cannot earn excess profits, and so there appears to be no incentive to enter. A policy of setting the price p_m^* and output q_m^* , together with the 'declared intention' of maintaining this output in the event of entry taking place and allowing the entrant's output to 'spoil the market', appears to offer to the monopolist maximum profit consistent with complete exclusion of new entry.

The central aspect of the analysis is the entrant's belief that the monopolist will continue to produce the output q_m^* in the event of her entry. It is in fact this output, rather than the price as such, which is the crucial feature of the entry-excluding strategy – the price p_m^* simply follows from the need to sell this output on the market – and the theory could perhaps be more aptly described as one of 'limit output'.

Credible threats and entry deterrence

Our analysis of the limit price did not specify how the entrant could be made to believe that the monopolist would in fact maintain the output q_m^* . This is a severe limitation, because it is not hard to construct an argument that suggests the potential entrant would not believe the threat. The entrant could reason that if she enters, say, at a scale q_e^b total output $q_m^* + q_e^b$ would drive price down to p_b which is assumed below the incumbent monopoly's minimum average cost (not shown in the figure). Then, both firms make losses for as long as the monopoly produces q_m^* . The entrant may believe that, once entry has taken place, the incumbent will realize that such continued losses are pointless, and the firms will move to an equilibrium in which both make positive profits. So, the entrant moves into the market. The monopoly's threat has not been credible.

The weakness of the limit pricing analysis is that it ignores the question of the *strategic interdependence* between the incumbent firm and the potential entrant. The two firms are involved in a *game*, and the theory of games tells us that the credibility of threats cannot simply be assumed, but instead must be explicitly analysed. In the rest of this section we examine some aspects of the game-theoretic approach to entry into a monopolized market.

Fig. 11.8 shows two *game trees*, (a) corresponding to a 'strong monopolist' M_s , and (b) to a 'weak monopolist' M_w . Each tree should be read as follows. The circled E indicates that the entrant must make a choice of moves, and she may choose to move *in* or to stay *out*. The incumbent monopoly must decide what to do given the entrant's choice. If this

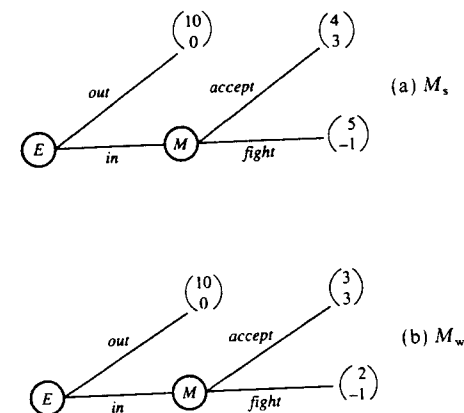


Fig. 11.8

has been *in*, then the monopolist must decide whether to *fight* (as in the above limit pricing analysis) or to *accept* the entry. The resulting profit payoffs to the two firms are shown in the brackets, with the monopolist's above and the entrant's below. If the entrant stays out, then the monopolist simply has a payoff consisting of his maximum monopoly profit. The difference between the two game trees is in the pattern of payoffs.

The game tree is assumed to be known to the players in each case, each player knows that the other knows it, and so on, an assumption referred to as the *common knowledge assumption*. Consider how E will reason in case (a), where she is playing against M_s . If she chooses *in*, then M_s will certainly choose *fight*, because his payoff at 5 is greater than his payoff if he accommodates to the new entry, at 4. Thus E will believe any threat M_s may make that he will *fight*, and choose *out*. In other words, a threat is *credible* if it is clearly in a player's interest to carry it out when called upon to do so, and this is why we call case (a) the 'strong monopolist' case.

In (b), such a threat is clearly not credible. If E chooses *in*, M_w does better by accepting rather than fighting, a payoff of 3 as opposed to 2, and so E will enter the market because her payoff from the sequence (*in*, *accept*) is greater than that from staying out. Note that we must assume that M_w will act rationally following E 's choice of *in* – as the game is constructed, *accept* is really the best thing for him to do in that case.

The difference between the 'weak' and 'strong' monopolist clearly lies in the relation between the payoffs to fighting and accepting entry. The payoffs to fighting entry could be thought of as being determined by the relative costs of the incumbent and entrant – the strong monopolist has the benefit of higher entry barriers, or some other advantage which makes a price war less damaging. We construct these payoffs in a particular example below. The payoffs to both firms following acceptance of entry depend on the exact nature of the *duopoly equilibrium*. For the moment we simply assume that the firms can predict the payoffs to which this will give rise.

This simple game analysis suggests another weakness of the previous limit pricing theory.

There is no need for the incumbent monopolist to forego short-run profits by setting the limit price p_m^* . If a potential entrant appears, then the strong monopolist can go on earning those profits because, if the potential entrant is rational and well-informed, she knows that it is not worth entering, while the weak monopolist knows that if an entrant appears the days of high profit are over in any case. Note also that the game analysis suggests that we would never actually observe *contested entry*, i.e. entry followed by a defensive fight whether successful or unsuccessful. Since such fights are observed, we may question whether the players are assumed to have too much information.

If we assume that the entrant is imperfectly informed about the monopolist's true cost structure in relation to her own, then it is easy to explain contested entry. In terms of the game analysis, we could interpret this imperfect information as implying that the entrant does not know for sure which type of monopolist she faces, M_s or M_w , but that she can assign prior probabilities to these. Let us assume the entrant is *risk neutral*; that is, she evaluates any risky decision in terms of the expected value of the profit associated with it.¹ We can calculate a critical probability \hat{p} that the entrant may assign to the event that she faces M_s , such that if the probability she actually assigns is $p < \hat{p}$ she will enter the market.

Thus, if p is the probability that the monopolist is strong, and if the entrant moves into the market, she receives a payoff of -2 with probability p and a payoff of 3 with probability $1 - p$, since M_s will fight and M_w will accept. The expected value of profit for the choice of *in* is therefore $-2p + 3(1 - p) = 3 - 5p$. Since the entrant is risk neutral she will enter if this is positive and stay out if this is negative, and so the critical value \hat{p} satisfies

$$3 - 5\hat{p} = 0 \Rightarrow \hat{p} = 0.6$$

If the entrant believes that the probability that the monopolist is strong is anything below 0.6 she will enter. Then, if she is wrong, we will observe contested entry and the subsequent exit of the would-be entrant. The greater the positive payoff from sharing the market with a weak monopolist relative to the negative payoff from running into a strong monopolist, the larger is \hat{p} and the greater the range of prior probability beliefs of the entrant that are consistent with entry.

Let us now concentrate on the case that the monopolist is known to be weak. It seems impossible to deny the logic of the argument that any threat to resist entry in this case is simply not credible, since once entry has taken place the incumbent does better by not fighting. One rationale for the intuitive feeling that this somehow is too simple is that it may pay even the weak monopolist to fight entry on one occasion, if this will establish a reputation for combativeness and so discourage entry at any time in the future. This changes the specification of the game, from a *one-shot game* to a *repeated game*, the *constituent game* of which we can take as given in Fig. 11.8(b).

Much depends on the firm's *discount rate*, that is, the rate at which it discounts future profits to a present value.² Even more important is whether the game is to be repeated a *finite* or *infinite* number of times. To emphasize the latter point, we assume that the firm does not discount the future, and we take Selten's (1978) Chain Store Paradox. The paradox is that even though it appears in the monopoly's interest to fight entry, it will accept entry from the very beginning.

Suppose a supermarket chain has stores in ten cities, in each of which it enjoys a local monopoly. In each city there is a local potential entrant, and the situation in each city is

represented in Fig. 11.8(b). The potential entrants act sequentially, so we number cities from 1 to 10 in the order in which entry will be attempted. Consider city 1. We might feel that the monopoly ought to fight any entrant into that market, because, although it sacrifices a payoff of 1 as compared to the accept strategy, if this then deters entry in the remaining 9 cities, it will gain $90 - 27 = 63$. But consider the last market to be defended, in city 10. Since there is no subsequent threat of entry to be deterred, the earlier logic of the one-shot game holds and the monopoly does better by accepting entry. All participants in the game, including the other nine potential entrants, can work this out. Take now city 9. Since the monopoly's decision in city 10 is to accept entry, it gains nothing in city 9 by fighting – the entrant in city 10 would not be deterred because she knows that once she enters, it is in the monopolist's interest to accept. Thus in city 9 the entrant will move in and the monopoly's best choice is to accept. But this argument can then be repeated for each market right back to city 1. Thus entry takes place in all markets.

Another way of interpreting this argument is the following. Everyone knows that the monopolist is weak. If he were to fight entry in the first city this would not change anyone's knowledge that in fact he is a weak monopolist: the entrant in city 10 would still believe that her entry will be accepted, and so will the entrant in city 9, and so on down the line. The incumbent will simply then be faced with a sequence of entries, each of which he would do better to accept than fight.

Again, this conclusion rests on the completeness of the information available to all the entrants. We could relax this by supposing that each entrant is uncertain whether the incumbent monopoly is weak or strong, with prior probabilities attached to these events. It can then be shown that under some conditions it *will* pay the weak monopolist to fight entry at earlier stages in the sequence, in order to create a reputation for being strong. He can exploit the uncertainty in the entrants' mind about his true type. In the later stages he will accept entry because the benefits of deterring future entry fall relative to the costs of fighting entry.

Capacity as credible precommitment

Returning to the case of the one-shot game, clearly the problem for the weak monopolist is the lack of credibility of his threat to fight entry – his bluff can too easily be called. A way to avoid this problem would be to find some means of *credible precommitment* to fighting, so that the entrant would expect her entry to be contested. One possibility, for example, might be for the monopolist to appoint a manager to take the firm's decisions, and to draw up a binding contract under which the manager is paid as a function of the firm's market share – perhaps with a sharp fall in pay for anything less than 100 per cent of the market! The manager then has a powerful incentive to fight entry, and provided the owner of the firm can credibly commit not to interfere with management, the entrant will not enter. (This is an example of a case in which a principal can do better by appointing an agent with different preferences to his own).

A different type of precommitment underlies the *Spence–Dixit model of entry-deterrence*. Suppose that the incumbent is able to commit to a particular level of capacity before entry takes place. For this to represent a credible commitment, we must assume that once installed, the capacity cannot be sold off at the market price for capacity – it represents

a *sunk cost*. For this model it is also important to assume that if the potential entrant comes into the market, the incumbent can expand capacity as the entrant is installing hers. Thus there is an asymmetry in the costs to the incumbent of expanding and contracting capacity. A second asymmetry is that the entrant must incur a fixed cost upon entry, although there is a constant marginal cost per unit of capacity output. This implies a falling average total cost curve with respect to the entrant's output, and so captures the idea of economies of scale for the entrant. The incumbent incurred his fixed cost in the past and so this is no longer relevant to his decisions. The importance of this model is that it brings out clearly the way in which the incumbent can use capacity precommitment either to forestall entry completely, or to improve his profitability should entry take place. For concreteness we analyse the model in the form of a specific numerical example, but in Exercise 11D a number of questions explore the effects of varying the assumptions.

The inverse market demand function is given by

$$p = 100 - x \quad [\text{D.1}]$$

where x is total market output. The incumbent is denoted firm 1, the entrant firm 2, and their outputs are x_1 and x_2 respectively. Capacity is denoted k_i , $i = 1, 2$, and each firm can install capacity at a constant cost of 30 per unit. Increases in output below capacity have a zero marginal cost, but output cannot exceed capacity at any cost. The entrant, if she decides to enter, has to pay a fixed cost F . Thus the firms' cost functions and capacity constraints are

$$C_1 = 30k_1, \quad x_1 \leq k_1; \quad C_2 = F + 30k_2, \quad x_2 \leq k_2 \quad [\text{D.2}]$$

We consider first the situation in which firms choose capacities simultaneously, and assume that their choices are a *Cournot-Nash equilibrium*. The rationale for this equilibrium concept is discussed fully in the next chapter. Here we simply note that it requires that the firms' choices be *mutually best responses*: firm 1's output and capacity must maximize its profit given firm 2's output and capacity, which in turn must maximize firm 2's profit given firm 1's choices. Formally, we find the equilibrium by deriving each firm's *reaction function*. For given x_2 , solve for firm 1 the problem

$$\max_{k_1, x_1} \pi_1 = [100 - (x_1 + x_2)]x_1 - 30k_1 \quad \text{s.t. } x_1 \leq k_1 \quad [\text{D.3}]$$

giving the solution

$$x_1 = k_1 = 35 - 0.5x_2 \quad [\text{D.4}]$$

which is 1's reaction function showing its optimal decision as a function of the given x_2 . Likewise for firm 2 solve

$$\max_{k_2, x_2} \pi_2 = [100 - (x_1 + x_2)]x_2 - 30k_2 - F \quad \text{s.t. } x_2 \leq k_2 \quad [\text{D.5}]$$

giving

$$x_2 = k_2 = 35 - 0.5x_1 \quad [\text{D.6}]$$

Note that it never pays to set output below capacity, and that the reaction functions are symmetrical, despite the presence of the fixed cost to firm 2. This latter does, however, play an important role, as we shall soon see.

The Cournot-Nash equilibrium is the pair of capacities k_i^c which satisfy the reaction functions simultaneously, so that we have

$$k_1^c = k_2^c = 23.33 \quad [\text{D.7}]$$

Then, market price is $(100 - 46.66) = 53.33$, and profits are

$$\pi_1^c = 544.44 \quad \pi_2^c = 544.44 - F \quad [\text{D.8}]$$

The reaction functions [D.4] and [D.6] are labelled R_1 and R_2 in Fig. 11.9 (ignore \bar{R}_1 for the moment) and the above Cournot-Nash equilibrium is also illustrated. The intercept of R_1 on the x_1 -axis is firm 1's monopoly output, since it corresponds to $k_2 = x_2 = 0$. As we move along R_1 leftwards firm 1's profit is falling since x_2 is increasing; and similarly as we move along R_2 rightwards firm 2's profit is falling since x_1 is increasing. Finally, note that if $F \geq 544.44$, the potential entrant cannot make a profit at the post-entry equilibrium, and so will not enter. Thus entry-deterrence would be irrelevant and the incumbent could set monopoly output and price. In what follows therefore we assume throughout that $F < 544.44$.

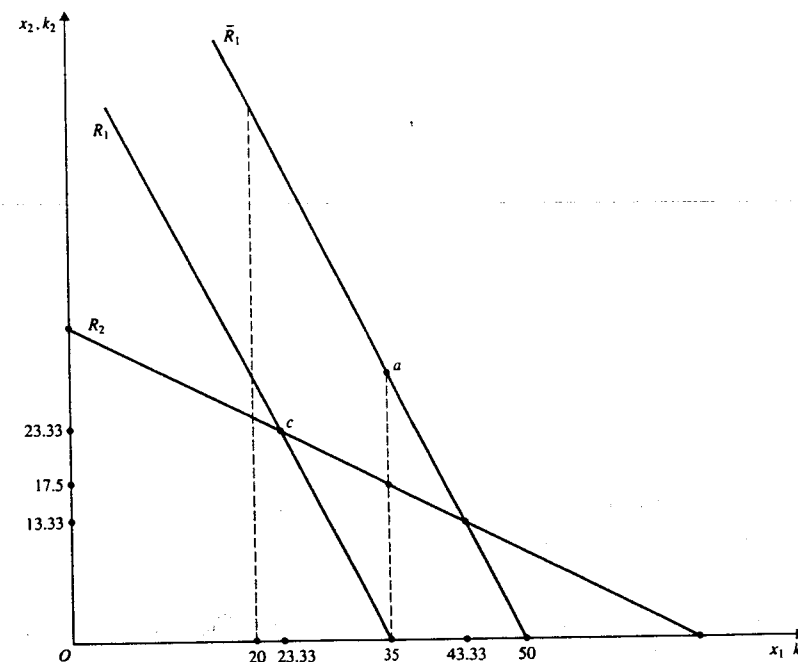


Fig. 11.9

Now we introduce the possibility that firm 1 can commit itself to a capacity level *before* the entrant appears. We emphasize the assumption that it cannot reduce its costs by reducing capacity *after* the entrant appears – committing itself to a capacity implies that it then has a fixed cost of $30\bar{k}_1$ and zero variable costs. In this case we find its output reaction function for a *given* commitment to capacity, \bar{k}_1 by solving

$$\max_{x_1} \pi_1 = [100 - (x_1 + x_2)]x_1 \quad \text{s.t. } x_1 \leq \bar{k}_1 \quad [\text{D.9}]$$

yielding

$$\begin{aligned} x_1 &= 50 - 0.5x_2 & \text{when } x_1 < \bar{k}_1 \\ x_1 &= \bar{k}_1 & \text{otherwise} \end{aligned} \quad [\text{D.10}]$$

This is illustrated as \bar{R}_1 in Fig. 11.9, for three assumed values of \bar{k}_1 , at $\bar{k}_1 = 20$, $\bar{k}_1 = 35$, and $\bar{k}_1 = 50$. Note that if $\bar{k}_1 = 50$, the reaction function \bar{R}_1 intersects firm 2's reaction function at the point (43.33, 13.33). The significance of this point is that if firm 1 had installed a capacity of 50 (or *any* $k_1 \geq 43.33$), then when the entrant appears and chooses a capacity and output, while firm 1 simply chooses an output (the marginal cost of which is *zero* up to 50), then this intersection point would be the Cournot–Nash equilibrium of this game, since it is the point at which the firms are making mutually best responses. On the other hand, if firm 1 had initially installed capacity $\bar{k}_1 < 43.33$, this point would not be feasible in the post-entry game.

We now examine firm 1's optimal choice of a prior capacity \bar{k}_1 . First, we show that it will never choose $\bar{k}_1 < 23.33$, or $\bar{k}_1 > 43.33$. Thus, the two reaction function intersection points define an interval [23.33, 43.33] within which firm 1 will choose its capacity.

Note first that if it chooses $\bar{k}_1 > 43.33$, the post-entry equilibrium will be at $x_1 = 43.33$. But then any excess of capacity over 43.33 is wasted, and since it costs 30 per unit to install, will not be chosen.

The argument that the incumbent will not choose $\bar{k}_1 < 23.33$ rests on the assumption that he can always increase his capacity in the post-entry period, just as the entrant is installing hers. It follows that choice of $\bar{k}_1 < 23.33$ is not a credible threat to the entrant. If the entrant chooses $k_2 = 23.33$, firm 1's best response is to expand capacity to 23.33 (moving along its *lower* reaction curve since it is having to buy capacity) and so setting $\bar{k}_1 < 23.33$ is pointless. The incumbent can commit to a minimum level of capacity \bar{k}_1 but cannot commit not to install additional capacity after entry.

The incumbent therefore will choose his most profitable minimum capacity commitment in the interval [23.33, 43.33]. Its level will depend on the value of the entrant's fixed cost. We can distinguish two cases:

1. At $k_2 = 13.33$, the entrant would make a profit. Since at the point (43.33, 13.33), we have that $\pi_2 = 177.69 - F$, this case corresponds to

$$0 < F < 177.69$$

It then follows that for any \bar{k}_1 in the interval [23.33, 43.33] the entrant can make a profit (recall the entrant's profit increases as we move leftward along R_2) and so entry cannot be deterred. All the incumbent can do is to choose \bar{k}_1 so as to make the post-entry situation

as profitable to himself as possible. Now for any \bar{k}_1 he chooses, the entrant will choose a point on R_2 . Thus, the incumbent can find his optimal capacity and output by using [D.6] to substitute for x_2 in his profit function and solving

$$\max_{x_1, \bar{k}_1} [100 - (x_1 + 35 - 0.5x_1)]x_1 - 30\bar{k}_1 \quad \text{s.t. } x_1 \leq \bar{k}_1 \quad [\text{D.11}]$$

This yields $x_1^* = \bar{k}_1^* = 35$. If the incumbent commits to this capacity level, the entrant's most profitable output and capacity are $x_2^* = k_2^* = 17.5$. Thus market price will be 57.5 and the firms' profits will be

$$\pi_1^* = 962.5; \quad \pi_2^* = 481.25 - F \quad [\text{D.12}]$$

This solution, in which the incumbent in effect maximizes his profit subject to the entrant's reaction function as a constraint, is known as the *Stackelberg leadership solution* and is more fully discussed in the next chapter. It is entirely an accident of the parameter values in this example that this solution happens to be at the incumbent's monopoly output level of 35. In general, the Stackelberg solution could be above or below the monopoly output. It is, however, not an accident that $x_1^* = \bar{k}_1^*$: it would not pay the incumbent to install capacity he would not use. The incumbent would never expand capacity post-entry, while the assumption that capacity, once installed, cannot be sold off makes it impossible for the incumbent to move back along R_1 in response to any higher capacity choice than 17.5 by the entrant. The fact that capital cost is a sunk cost enables the incumbent credibly to commit to an output and capacity of 35. In effect this precommitment makes the reaction function of firm 1 the portion of \bar{R}_1 from 100 to a , and the vertical dashed line from a to 35. Then, (35, 17.5) is an intersection of the reaction functions and hence a Cournot–Nash equilibrium.

2. The other possibility is that the entrant's fixed cost lies in the interval

$$177.69 \leq F < 544.44$$

implying that his profits become zero at some point on R_2 between (23.33, 23.33) and (43.33, 13.33). We can think of this point as being fixed by its x_1 -coordinate. Then we can solve for this x_1 -coordinate, as a function of F , by noting that it satisfies the condition

$$\pi_2 = [100 - (x_1 + x_2)]x_2 - 30x_2 - F = 0 \quad [\text{D.13}]$$

Substituting the firm 2 reaction function $x_2 = 35 - 0.5x_1$ into this condition yields the function

$$\hat{x}_1 = 70 - 2F^{1/2} \quad \text{for } 177.69 \leq F < 544.44 \quad [\text{D.14}]$$

(strictly, [D.13] gives a quadratic and therefore two solutions for \hat{x}_1 , but only the root given in [D.14] is feasible). The significance of \hat{x}_1 is that if the incumbent sets $\bar{k}_1 \geq \hat{x}_1$, then entry will be deterred, since the entrant's best response to this will yield her no profit, given the level of her fixed cost. Again, the irreversibility of the investment makes this precommitment a credible means of deterring entry. Is it a profit-maximizing strategy for the incumbent to choose $k_1 = \hat{x}_1$ and deter entry? We can distinguish two sub-cases:

- (a) $23.33 < \hat{x}_1 \leq 35$, corresponding to $306.25 \leq F < 544.44$.

In this case, by setting $\bar{k}_1 = x_1 = 35$ the incumbent deters entry and maximizes monopoly profit.

- (b) $35 < \hat{x}_1 \leq 43.33$, corresponding to $177.69 \leq F < 306.25$.

The optimal capacity choice in this case requires a comparison of the profit to be made by permitting entry, on the one hand, or by setting $\bar{k}_1 = \hat{x}_1$ and so preventing entry, on the other. We already saw that if entry is going to take place, the best output for the monopolist is his Stackelberg output of 35, yielding post-entry profit of 612.5. On the other hand, if the incumbent preserves his monopoly at any output over the interval $[35, 43.33]$, his profit lies in the interval $[1155.61, 1225]$. Clearly then it pays him always to deter entry and to set $\bar{k}_1 = \hat{x}_1$ (again, note that as long as capacity is costly we would never have $\bar{k}_1 > \hat{x}_1$, since such excess capacity is unnecessary).

In general, a third subcase is possible. Over an upper part of the interval $[35, 43.33]$ it might have been the case that the monopoly's profit from setting $\bar{k}_1 = \hat{x}_1$ was below that from setting \bar{k}_1 at the Stackelberg point. Then the latter would be optimal. For the parameter values in this example, however, that never happens.

In Table 11.1 we summarize the various cases just analysed. They are defined in terms of the entrant's fixed cost. Only in the lowest range of F -values does entry take place, while in the simultaneous game entry occurs for $F < 544.44$. Even when entry occurs, we see that the incumbent's output and profit are both higher than in the simultaneous game, in which the Cournot–Nash equilibrium gave the incumbent a profit of 544.44. By being able to precommit to capacity the incumbent has a higher reaction function in the post-entry game and so secures a more favourable equilibrium. In the other cases precommitment allows the incumbent actually to prevent entry and earn even higher profit, and in one range the incumbent can earn monopoly profit.

Table 11.1 Possible equilibrium in capacity precommitment game

Fixed cost	Entrant chooses	Incumbent chooses	Incumbent profit
$0 < F < 177.69$	In, $k_2 = 17.5$	$x_1 = \bar{k}_1 = 35$	962.5
$177.69 \leq F < 306.25$	Out, $k_2 = 0$	$x_1 = \bar{k}_1 = 70 - 2F^{1/2}$	$(1155.44, 1225]$
$306.25 \leq F < 544.44$	Out, $k_2 = 0$	$x_1 = \bar{k}_1 = 35$	1225
$544.44 \leq F$	Out, $k_2 = 0$	$x_1 = \bar{k}_1 = 35$	1225

Exercise 11D

1. Show that the qualitative nature of the analysis of the capacity precommitment (Spence–Dixit) model is unaffected if we assume a constant positive marginal cost of output for each firm.
2. In the example of the Spence–Dixit model given in this section, take the entrant's fixed cost $F = 300$, and analyse the implications of taking different possible values of the marginal capacity cost c (in the example it was set at 30).

3. In the Spence–Dixit model, what would be the consequence of assuming that the incumbent *could* sell off units of capacity, post-entry, but at a price less than the cost of installing new capacity?
4. What happens in the Spence–Dixit model if the entrant has zero fixed cost?

Notes

1. See Chapter 19 for a full discussion of the meaning of risk neutrality.
2. See Chapter 15 for a full discussion of present values. Here, we simply make use of the fact that the present value of any infinite stream of constant payments a , beginning in one period's time, is simply a/r , where r is the per-period interest rate.

References and further reading

The classic paper on monopoly is:

- A. P. Lerner. 'The concept of monopoly and the measurement of monopoly power', *Review of Economic Studies*, 1, 1934.

For a very comprehensive treatment of all the topics discussed in this chapter, and much more, see:

- J. Tirole. *The Theory of Industrial Organization*, The MIT Press, Cambridge, Mass., 1988.

On price discrimination, see:

- H. Varian. 'Price discrimination', ch. 10 of R. Schmalensee and R. D. Willig (eds), *Handbook of Industrial Organization*, North-Holland, Amsterdam, 1989.

On entry see:

- R. J. Gilbert. 'Mobility barriers and the value of incumbency', ch. 8 of R. Schmalensee and R. D. Willig (eds), *Handbook of Industrial Organization*, North-Holland, Amsterdam, 1989.

The classic article on the Chain-Store Paradox is:

- R. Selten. 'The Chain-Store Paradox', *Theory and Decision*, 9, 1978, 127–59.

An excellent exposition of the capacity pre-commitment model in more general terms than the example given here, is:

- A. Dixit. 'The role of investment in entry deterrence', *Economic Journal*, 90, 1980, 95–106.

CHAPTER 12

Oligopoly

A. Introduction

If a firm believes that the outcome of its decision depends significantly on the decisions taken by one or more other identifiable sellers, then we have the market situation known as oligopoly. It is usual to define oligopoly as a market with 'few sellers' (and indeed the word 'oligopoly' means this), but, as we argued in the previous chapter, a definition in terms of the number of sellers in a market is not without ambiguity. Since the essence of the situation is the nature of the competitive relationships between sellers, it is best to make this the basis of the definition. Nevertheless, loosely and intuitively we always think of oligopoly as 'competition among the few'.

We assume that a firm in this situation of close interdependence of decision-taking will seek to maximize profit. The problem it faces is to assign a profit outcome to each decision alternative (e.g. production plan), in order to rank them and find the optimum. Each firm is necessarily involved in reasoning: 'if I choose *A*, and he chooses *B*, then I get *X*, while if I choose *C*, and he chooses *D* then I get *Y*...', and so on. In this example, the competitor's reactions, *B*, *D*, ... could take one of a number of forms, and so the firm in question must try to reason out what the response will be. Before he can come to a ranking of alternative decisions, he must take some view of the actions of each of his competitors. The theory of oligopoly is concerned with understanding and predicting the decisions of sellers in such situations of 'strategic interdependence', i.e. of interactions of reasoning and decision-taking among sellers.

A natural way to proceed would seem to be to formulate a particular hypothesis about the nature of the competitive reactions which each firm expects and use this to find an equilibrium solution. The hypothesis allows us to say that each decision-taker will associate with his decision *A* some specific response *B*, and with his decision *C* some specific response *D*, and so on. Then, by using the basic analytical framework of cost and demand curves, we arrive at a precise prediction of the market equilibrium. This approach was indeed the one first adopted by economists. However, there are several hypotheses about reaction patterns which are possible, each with a different associated equilibrium solution. We then have several possible theories with different solutions. This in itself need not be a serious cause for concern, since we could presumably use empirical evidence to distinguish among

the various hypotheses and find that which appeared to be the best representation of sellers' beliefs about reaction patterns in any given market.

The application of game theory to oligopoly theory has led to a fundamental reinterpretation of these models. The game-theoretic approach does not allow choice of an arbitrary, even if plausible, pattern of reactions. Rather, beliefs about the actions of a competitor have to be shown to follow from rational calculation by the firm concerned. Though, as we see below, several of the traditional oligopoly models still hold a central place, the emphasis on rational calculation has led to a more careful definition of the types of market situation for which they can be expected to hold, and a deeper understanding of the models themselves.

A further criticism of the 'reaction pattern' approach is that it ignores the possibility of explicit communication and cooperation among sellers. It appears to assume that the sellers remain 'at arm's length', guessing at each other's likely reactions, whereas an obvious possibility is that they would at least consult and quite conceivably cooperate. To quote a famous passage from Adam Smith's *Wealth of Nations*: 'People of the same trade seldom meet together, even for merriment or diversion, but the conversation ends in a conspiracy against the public, or in some contrivance to raise prices. It is impossible indeed to prevent such meetings, by any law which either could be executed, or would be consistent with liberty and justice' (vol. 1, p. 117, Everyman edition). The modern phenomenon of the expense account lunch fits Smith's description of a meeting 'for merriment or diversion' quite accurately. Admitting the possibilities of communication and cooperation changes the focus of the analysis. Instead of constructing hypotheses about expected reaction patterns and examining their consequences, we are interested in answers to the questions:

- Under what conditions will the sellers agree to cooperate in their decisions?
- If they agree to cooperate, what price and output policies will result?
- Will their cooperative agreement be *stable*, in the sense of being maintained over time in the face of changing circumstances, and, in particular, are there forces making for a breakdown of the agreement?

Whether profit-maximizing firms will agree to cooperate depends crucially on the number of times the market situation is repeated. The traditional oligopoly models implicitly treat the market situation as a *one-shot game*: the firms produce and sell outputs just once. It turns out that in this case it is very difficult to rationalize collusive behaviour. If on the other hand we view the market situation as being repeated (possibly infinitely) many times, it becomes quite easy to explain collusive behaviour, and the difficulty becomes that of explaining the precise prices and quantities that will be chosen.

In this chapter we do not attempt to survey the enormous literature on oligopoly. Instead we use game theory to analyse a small number of the more important oligopoly models. In doing so we introduce the reader to several important concepts from non-cooperative game theory, concepts which we use again in later chapters. In section B firms play a one shot duopoly game and we use the notion of a Nash equilibrium in pure and mixed strategies to consider four alternative models suggested by Cournot, Stackelberg, Bertrand and Edgeworth. In sections C and D firms interact repeatedly, leading us to introduce

several ideas from the theory of dynamic games (subgame perfection, the Folk Theorem, and renegotiation proofness) to examine whether collusive outcomes can be sustained by threats of future retaliation.

B. One-shot games

In this chapter we work in terms of a very specific model. The advantage is that the central results can be shown very simply and clearly. The disadvantage is that it is not always clear how or whether these results generalize: the general issues of existence, uniqueness and stability of equilibria are not dealt with. Some remarks are made on these issues, but the interested reader is directed to the more specialized references at the end of the chapter for a fuller treatment.

We assume there are just two firms, with total cost functions

$$C_i = c_i q_i \quad i = 1, 2 \quad c_i > 0 \quad [\text{B.1}]$$

so that the firms have constant marginal costs. The outputs of the firms may or may not be homogeneous. If they are homogeneous then $c_1 = c_2 = c$. The inverse demand function for output of firm i is assumed to be

$$p_i = \alpha_i - \beta_i q_i - \gamma q_j \quad i, j = 1, 2 \quad i \neq j \quad [\text{B.2}]$$

where $\gamma > 0$. The goods are therefore substitutes: an expansion of firm j 's output (corresponding to a fall in its price) pushes down the demand and revenue functions of firm i . The symmetry of cross-partial $\partial p_i / \partial q_j$ is useful, but recall from Chapter 4 that we would not in general expect it to exist for Marshallian demand functions, so [B.2] involves a substantive restriction on the nature of demands for these goods. We assume $\alpha_i > c_i$ so that the markets are active.

If the firms' outputs are homogeneous, then

$$\alpha_1 = \alpha_2 = \alpha \quad \text{and} \quad \gamma = \beta_1 = \beta_2 \quad [\text{B.3}]$$

and so the outputs must sell at identical prices determined by the sum of the firms' outputs – we have in effect only one demand function, $p = \alpha - \gamma(q_1 + q_2)$. The firms' profit functions, as functions of outputs, are

$$\pi_i(q_1, q_2) = p_i q_i - c_i q_i = (\alpha_i - c_i - \gamma q_j) q_i - \beta_i q_i^2 \quad i, j = 1, 2 \quad i \neq j \quad [\text{B.4}]$$

If the firms' outputs are not homogeneous, we can use the inverse demand functions in [B.2] to get the demand functions

$$q_i = q_i(p_1, p_2) = a_i - b_i p_j + \phi p_i \quad i, j = 1, 2 \quad i \neq j \quad [\text{B.5}]$$

where a_i, b_i and ϕ are all positive. (See Question 1, Exercise 12B for the precise definitions of these parameters and the implied restrictions on the α_i, β_i and γ .) Note that the individual demand functions cannot be expressed in this way if the outputs are homogeneous. We can also think of profit as a function of prices

$$\pi_i(q_1(p_1, p_2), q_2(p_1, p_2)) \quad i = 1, 2 \quad [\text{B.6}]$$

Note, finally from [B.4] that the profit function π_i is strictly concave in q_i for given q_j , with a maximum at

$$q_i = \frac{\alpha_i - c_i - \gamma q_j}{2\beta_i} \quad i, j = 1, 2 \quad i \neq j \quad [\text{B.7}]$$

while it is linear and decreasing in q_j for given q_i . Its Hessian determinant is

$$\begin{vmatrix} -2\beta_i & -\gamma \\ -\gamma & 0 \end{vmatrix} = -\gamma^2 < 0 \quad [\text{B.8}]$$

which implies that the function is *not concave* (recall Chapter 2, section 1). However, it can be shown that the function is strictly quasi-concave (see Question 7, Exercise 2B).

We now consider some oligopoly models in the context of this rather well behaved example.

1. The Cournot model

Assume the market operates as follows. Each firm must decide, without consulting the other, what output it will produce. The firms simultaneously put their outputs on the market. Prices then move to the levels that clear the market, and the firms receive the resulting profits. What outputs will they choose?

First, note that [B.7] gives each firm important information. Given any output q_j that firm i expects the other to produce, its *best response* is to produce the q_i given by [B.7]. Accordingly [B.7] defines the firms' *best response functions*,

$$q_i = A_i - B_i q_j \quad i, j = 1, 2 \quad i \neq j \quad [\text{B.9}]$$

where $A_i \equiv (\alpha_i - c_i)/2\beta_i$; $B_i \equiv \gamma/2\beta_i$.

Examples of these functions are graphed in Fig. 12.1. The slopes are negative because increases in output q_j reduce firm i 's profit maximizing output.

The intersection point of these two best response functions gives

$$q_i^c = \frac{A_i - A_j B_i}{1 - B_i B_j} \quad i, j = 1, 2 \quad i \neq j \quad [\text{B.10}]$$

by solving simultaneously the two equations defined by [B.9]. When the outputs are homogeneous, we have, using [B.3] in [B.10],

$$q_1^c = q_2^c = (\alpha - c)/3\gamma \quad [\text{B.11}]$$

where $-\gamma$ is the slope of the inverse market demand function.

In his original analysis of this model over 150 years ago, the French economist Augustin Cournot proposed this intersection point as the equilibrium outcome in the market. His argument went as follows. Suppose firm 1 puts output q_1^0 (in Fig. 12.1) onto the market. Then firm 2 will react by producing its profit maximizing output q_2^0 . But if firm 2 does this firm 1 will change its output to q_1^1 , which maximizes its profit when $q_2 = q_2^0$. This in turn induces firm 2 to change its output to q_2^1 . And so on. Since each firm reacts to the other's output by producing at a point on its best response function (or, in Cournot's terminology, *reaction curve*) the only possible equilibrium in the market, at which no

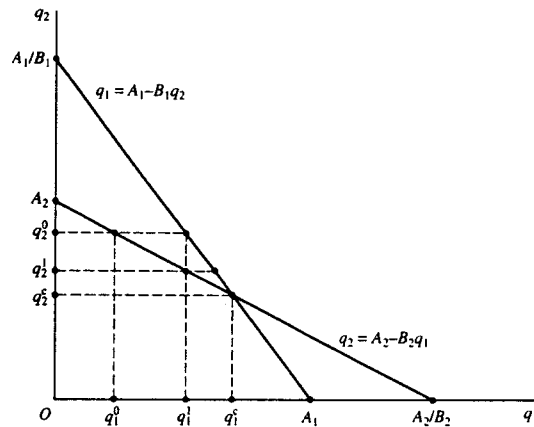


Fig. 12.1

further output changes will take place, is at the intersection point (q_1^*, q_2^*) . Here, neither firm wants to change its output given the other's output choice.

This argument is not very convincing. Notice that it is inconsistent with the one shot assumption since it requires that outputs are being chosen sequentially over a (possibly infinite) number of time periods. Each firm behaves myopically. Each expects that the other will keep its output constant, and this expectation about the other's reaction is held to, even though it is continually falsified. In section C we will see that rational firms may be able to do much better than the Cournot solution in multi-period situations. If the game is played only once, Cournot's process of sequential reactions to actual output choices cannot be used to rationalize the equilibrium at (q_1^*, q_2^*) .

The modern, game-theoretic treatment of this model provides a different rationale for the same equilibrium outcome. Each firm is assumed rationally to think through the consequences of its choices, in the knowledge that the other firm knows the situation and is also rationally thinking things through.¹ The outputs that the firms produce are then taken to be the *Nash equilibrium* output choices. The output pair (q_1^*, q_2^*) is a Nash equilibrium if

$$\pi_1(q_1^*, q_2^*) \geq \pi_1(q_1, q_2^*) \quad \text{and} \quad \pi_2(q_1^*, q_2^*) \geq \pi_2(q_1^*, q_2) \quad [\text{B.12}]$$

for all feasible outputs q_1, q_2 . That is, q_i^* must maximize i 's profit given that firm j is producing q_j^* , $i, j = 1, 2, i \neq j$. But clearly, the output pair (q_1^*, q_2^*) satisfies this definition and is moreover the only output pair that does so in this model. Thus the Nash equilibrium of this game is the Cournot equilibrium output pair, since it must be at the intersection of the firms' best response functions.

The argument underlying the use of the Nash equilibrium concept as the solution concept for games of this type is as follows. Suppose firm 2 thinks that firm 1 will produce output q_1^0 in Fig. 12.1. He can then calculate that q_2^0 is his best output on that assumption. But he then realizes that firm 1 can also work that out, and that if 1 thinks 2 will choose

q_2^0 , 1 will then want to produce q_1^1 . So 2 would be irrational to persist in believing that 1 will choose q_1^0 . Such an argument applies at every point except (q_1^*, q_2^*) . If 2 believes 1 will choose q_1^* , his best response is q_2^* , and if he thinks that 1 has also figured that out, he will not want to change his choice because q_1^* is 1's best response to q_2^* . The Nash equilibrium pair has the property that, if i knows j will choose q_j^* ($= q_j^*$), he will still wish to choose q_i^* ($= q_i^*$). The firms are then led to make this choice of outputs by going through the above reasoning process. The argument for applying the Nash equilibrium concept is that any non-Nash equilibrium point is open to the criticism that it would not be chosen by a player who believes that his opponent is as rational and well-informed as he is himself.

We can compare the Cournot-Nash equilibrium with two other possible output pairs. The first, denoted (\hat{q}_1, \hat{q}_2) corresponds to the 'perfectly competitive' solution² where price equals marginal cost:

$$p_i = \alpha_i - \beta_i \hat{q}_i - \gamma \hat{q}_j = c_i \quad i, j = 1, 2 \quad i \neq j \quad [\text{B.13}]$$

The second, denoted (q_1^m, q_2^m) , corresponds to the joint-profit maximization or monopoly solution, and so solves

$$\max_{q_1, q_2} \pi_1(q_1, q_2) + \pi_2(q_1, q_2) \quad [\text{B.14}]$$

with first-order conditions

$$\alpha_i - c_i - 2\gamma q_j - 2\beta_i q_i = 0 \quad i, j = 1, 2 \quad i \neq j \quad [\text{B.15}]$$

Table 12.1 shows the output solutions for each of these cases, for both differentiated and homogeneous outputs. Note that, because they have identical constant marginal costs, the firms' individual outputs are indeterminate in the competitive and monopolistic homogeneous output cases. It is left as an exercise (Question 2) to show that in the differentiated case

$$q_1^m < q_1^c < \hat{q}_1 \quad [\text{B.16}]$$

Table 12.1

Equilibrium outputs	Product differentiation	Homogeneous outputs
Cournot-Nash	$q_i^c = \frac{2\beta_j(\alpha_i - c_i) - \gamma(\alpha_j - c_j)}{4\beta_i\beta_j - \gamma^2}$	$q_i^c = \frac{\alpha - c}{3\gamma}$
Perfect competition	$\hat{q}_i = \frac{\beta_j(\alpha_i - c_i) - \gamma(\alpha_j - c_j)}{\beta_i\beta_j - \gamma^2}$	$\hat{q}_1 + \hat{q}_2 = \frac{\alpha - c}{\gamma}$
Monopoly	$q_i^m = \frac{\beta_j(\alpha_i - c_i) - \gamma(\alpha_j - c_j)}{2(\beta_i\beta_j - \gamma^2)}$	$q_1^m + q_2^m = \frac{\alpha - c}{2\gamma}$
Stackelberg (Firm 1 is leader Firm 2 is follower)	$q_1^s = \frac{2\beta_2(\alpha_1 - c_1) - \gamma(\alpha_2 - c_2)}{4\beta_1\beta_2 - 2\gamma^2}$ $q_2^s = \frac{2\beta_1(\alpha_2 - c_2) - \gamma(\alpha_1 - c_1) - \gamma A_2}{4\beta_1\beta_2 - 2\gamma^2}$	$q_1^s = \frac{\alpha - c}{2\gamma}$ $q_2^s = \frac{\alpha - c}{4\gamma}$

Thus outputs are lower and prices higher at the Cournot solution than under competition and so the firms earn positive profits; but prices and profits are not as high as when outputs are set to maximize joint profits. This is easily seen (for the sum of outputs) in the homogeneous case, since

$$(\alpha - c)/2\gamma < 2(\alpha - c)/3\gamma < (\alpha - c)/\gamma$$

If the firms really do want to maximize profits, why do they not collude and agree to set outputs q_i^m ? The answer is that, in this one-shot game, they will only do so if they can make a *binding commitment* to keep the agreed outputs. Otherwise, the attempt will fail because (q_1^m, q_2^m) is not a Nash equilibrium output pair.

Thus suppose the managers of the two firms meet, possibly for merriment and diversion, and agree to produce outputs q_i^m (or for that matter *any* outputs *other than* (q_1^m, q_2^m)). When they return to their firms and set about drawing up their production plans, the following thought will, if they are rational, occur to each of them. If the other firm is going to produce q_j^m , then i 's best response is not q_i^m but rather $q_i^R = A_i - B_i q_j^m$. It is an exercise to show that $q_i^R > q_i^m$. The 'R' stands for 'renege' – firm i is going to renege or cheat on the agreement, by producing a larger output than agreed, and thereby earning a still larger profit. But then, i will realize that j has also worked this out, and so producing q_j^m is even less of a good idea. The same process of reasoning described earlier will lead the two firms back to the Cournot–Nash equilibrium output pair.

How might they make a binding commitment? One possibility would be to draw up a legally binding contract, which provides for penalties of at least $\pi_i^R = \pi_i(q_i^R, q_j^m) - \pi_i(q_i^m, q_j^m)$ if i reneges. However, in many countries such an agreement would be illegal, so unenforceable, and therefore not binding. It is the possibility of punishing reneging by market sanctions, such as a price war, that makes the distinction between a one-shot and a repeated game such an important one, as we shall see in section C. In a one-shot game there is no next period in which to hold a price war. If they cannot find a way of making a binding commitment, the firms will not be able to agree to a more profitable output pair than (q_1^c, q_2^c) .

2. The Stackelberg model

Suppose that, instead of the firms making simultaneous output choices, firm 1 announces its output first and, once that announcement is made, the output cannot be changed. This makes firm 1 the *market leader*, and defines a model analysed by the German economist H. von Stackelberg. Firm 1 reasons that firm 2 will make the best response to its own announced output. There would be no point in firm 2's choosing its Cournot output, say, because firm 1 cannot then change its output from that announced. Thus, somehow, firm 1 is able to make a credible, binding commitment to an output level. What is firm 1's optimal output?

For any q_1 , firm 2 will choose q_2 on its best response function. Thus firm 1 chooses q_1 to maximize its profit subject to this constraint, i.e. it solves

$$\max_{q_1, q_2} \pi_1(q_1, q_2) \quad \text{s.t. } q_2 = A_2 - B_2 q_1 \quad [\text{B.17}]$$

The first order conditions

$$\alpha_1 - c_1 - \gamma q_2 - 2\beta_1 q_1 - \lambda B_2 = 0 \quad [\text{B.18}]$$

$$-\gamma q_1 - \lambda = 0 \quad [\text{B.19}]$$

$$q_2 = A_2 - B_2 q_1 \quad [\text{B.20}]$$

yield firm 1's Stackelberg output as

$$q_1^s = [2\beta_2(\alpha_1 - c_1) - \gamma(\alpha_2 - c_2)] / (4\beta_1\beta_2 - 2\gamma^2) \quad [\text{B.21}]$$

A comparison with the first row of Table 12.1 shows that $q_1^s > q_1^c$. Since this represents a rightward shift along firm 2's best response function, we must have $q_2^s < q_2^c$. In the homogeneous case

$$q_1^s = (\alpha - c)/2\gamma \quad [\text{B.22}]$$

giving the interesting (but special to the chosen functional form) result that $q_1^s = q_1^m + q_2^m$. Since $q_2^s > 0$, this tells us that again total profits are not maximized. Indeed, total profits are even lower than at the Cournot–Nash equilibrium, though firm 1's profit is greater, reflecting its *first mover advantage*.

Fig. 12.2 illustrates the Stackelberg outcome. The curves shown are the contours of firm 1's profit function $\pi_1(q_1, q_2)$. Firm 1's profits increase as we move *downward* in the figure, i.e. lower contours correspond to higher profits. Its best response function passes through the peaks of these contours because, for any q_2 , firm 1 finds the profit maximizing q_1 . The Stackelberg equilibrium is a point of tangency of a profit contour with 2's best response function, since [B.17] requires firm 1 to find the point on this function which is on the lowest possible profit contour. Note that the slope of a profit contour is

$$\frac{dq_2}{dq_1} = (\alpha_1 - c_1 - \gamma q_2 - 2\beta_1 q_1) / \gamma q_1$$

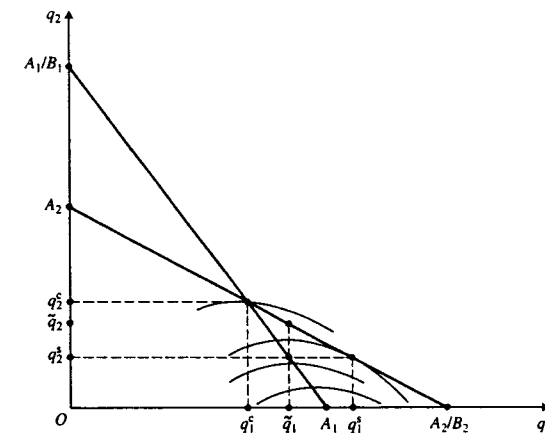


Fig. 12.2

and [B.18] and [B.19] then yield the condition

$$\frac{dq_2}{dq_1} = -B_2$$

Since the slope of firm 2's best response curve is $-B_2$, this implies the type of tangency point shown in Fig. 12.2.

The Stackelberg solution is also a Nash equilibrium of the game defined by the assumption that firm 1 credibly commits to an output level before firm 2 chooses its output level. firm 2 knows for sure that firm 1 will choose output q_1^* , then it will still want to choose output q_2^* ; and if firm 1 knows for sure that firm 2 will make its best response to 1's announcement, it will want to precommit to q_1^* and not some other level of output.

One might be tempted to argue at this point: but what about output \tilde{q}_1 in Fig. 12.2? Surely, if firm 1 knew firm 2 would produce q_2^* , then \tilde{q}_1 , on 1's best response function, yields higher profit than q_1^* . The answer is that this is a different game. If firm 1 can revise its output choice when it knows firm 2's output then \tilde{q}_1 would be the best response to q_2^* , but in that case we are back in the Cournot game and the outcome will be (q_1^c, q_2^c) . Firm 1 would never precommit to \tilde{q}_1 , because then firm 2 would choose its best response to this, at \tilde{q}_2 . Thus the Stackelberg outcome only makes sense as equilibrium when the leader is credibly committed to producing his announced output. (Chapter 11 gave an example of credible commitment where an incumbent firm installs capacity before a second firm tries to enter its market, and the cost of capacity is entirely a sunk cost.)

3. The Bertrand model

Up until now we have assumed that the firms choose outputs, with prices then being determined by the inverse demand functions [B.2]. In many oligopolistic markets firms appear to set prices, then sell what the market demands. In a monopoly it makes no difference whether we carry out the analysis in terms of prices or quantities as the choice variables. In his critical review of the book in which Cournot set out his oligopoly model, J. Bertrand showed that in the case of oligopoly the choice is crucial.

Suppose now that the firms choose prices simultaneously and independently then sell the outputs generated by the demand functions in [B.5]. For the moment assume that products are differentiated. What prices will they choose?

We again proceed by finding the Nash equilibrium of this game. First we require the best price-response functions, since a Nash equilibrium will be at their intersection. Thus, for given p_j , we solve

$$\max_{p_i} \pi_i = (p_i - c_i)q_i(p_i, p_j) = (p_i - c_i)(a_i - b_i p_i + \phi p_j)$$

$$i, j = 1, 2 \quad i \neq j \quad [\text{B.23}]$$

giving the best price-response function

$$p_i = \frac{a_i + c_i b_i}{2b_i} + \frac{\phi p_j}{2b_i} = \hat{A}_i + \hat{B}_i p_j \quad i, j = 1, 2 \quad i \neq j \quad [\text{B.24}]$$

where $\hat{A}_i \equiv (a_i + c_i b_i)/2b_i$, $\hat{B}_i \equiv \phi/2b_i$. The best price-response functions are linear with

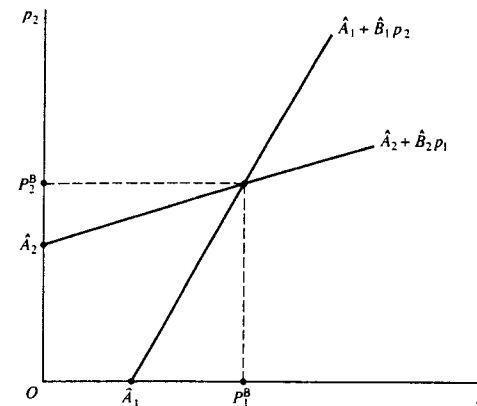


Fig. 12.3

positive slopes. As shown in Fig. 12.3, they intersect at the point (p_1^B, p_2^B) , where

$$p_i^B = \frac{\hat{A}_i + \hat{A}_j \hat{B}_i}{1 - \hat{B}_i \hat{B}_j} \quad i, j = 1, 2 \quad i \neq j \quad [\text{B.25}]$$

The Nash equilibrium in the Bertrand model is the pair of prices (p_1^B, p_2^B) . The rationalization of the equilibrium is on the same lines as in the Cournot model. No other pair of price choices has the property of mutual consistency. If player i begins by assuming that j will choose $p_j^0 \neq p_j^B$, and reasons things through, i must conclude that j will not choose p_j^0 , given that j has thought things through also.

It is an exercise (Question 5) to show that the best price response functions in [B.24] do intersect, and that they imply prices p_i^B which are greater than marginal cost c_i but less than the prices at the Cournot equilibrium. Thus, in the case of differentiated products, Bertrand prices and outputs are 'more competitive' than Cournot prices and outputs, but still generate excess profits.

It is in the case of homogeneous products that the consequences of the Bertrand analysis are the most striking. We can show that in this market the Nash equilibrium is at $p_1^B = p_2^B = c$, the competitive market outcome.

The Bertrand result is straightforward to establish. Suppose i expects j to set $p_j^0 > c$. Then i 's best response is to set $p_i^0 = p_j^0 - \varepsilon$, $\varepsilon > 0$, since this captures the entire market and for small enough ε gives i the highest profit possible. But i will then realise that j will have realized this, and so would plan $p_j^1 = p_i^0 - \varepsilon$, in which case i should set $p_i^1 = p_j^1 - \varepsilon$, ... and so on. Clearly in the end i cannot rationally believe that j will set $p_j > c$. But of course neither firm would set $p < c$, because this leads to losses. Thus $p_1 = p_2 = c$ is the only mutually consistent price pair in this market, and so it is the Nash equilibrium.

Bertrand intended this to be a *reductio ad absurdum*, to demonstrate the weakness of Cournot's approach. But both results are applications of the standard Nash equilibrium solution concept and there is nothing inherently unattractive about the idea that firms choose prices – quite the reverse. Thus the Bertrand outcome is a striking prediction from

a model that is in many respects reasonable. If we feel that the competitive outcome in a homogeneous market with two firms is somehow implausible, a leading candidate for revision is the one-shot game assumption.

It is possible that the extreme nature of Bertrand's result led to its relative neglect in economics until recently – the 'standard' oligopoly model was that of Cournot. Bertrand's model seemed to deprive oligopoly theory of much of its interest: in the homogeneous goods case, the move from one firm to two leads directly from monopoly to perfect competition! We next turn to a model which seeks to explore this further, with even more problematic results.

4. The Edgeworth model

Suppose that in the Bertrand homogeneous output model each firm has an exogenously given upper bound on capacity output, \bar{q}_i . In choosing prices the firms have to take account of the constraints $q_i \leq \bar{q}_i$, $i = 1, 2$.

For simplicity we assume,

$$\bar{q}_1 = \bar{q}_2 = \bar{q}$$

To simplify the notation we assume that the firms' identical marginal production cost is zero: $c = 0$. (As the reader could check all of the results below hold in the case in which $c > 0$.) Coupled with the assumption that the firms' capacities are exogenously determined, so that any costs of acquiring the capacity \bar{q} are fixed, the assumption that $c = 0$ implies that profit maximization is equivalent to revenue maximization.

Although the Edgeworth model is a capacity constrained price setting model, the best output response functions from the capacity constrained quantity setting game have an important role in the analysis. Because demand is homogeneous the inverse market demand function is given by $p = \alpha - \gamma(q_1 + q_2)$ and the best output response function for firm i is derived from the problem

$$\max_{q_i} \alpha q_i - \gamma(q_1 + q_2)q_i \quad \text{s.t. } q_i \leq \bar{q} \quad [\text{B.26}]$$

as

$$q_i = (\alpha/2\gamma) - q_j/2 \quad \text{for } q_i < \bar{q} \quad \text{and} \quad q_i = \bar{q} \quad \text{otherwise} \quad [\text{B.27}]$$

The best output response functions are graphed in Fig. 12.4. The Cournot–Nash outputs are $q_i^c = \alpha/3\gamma$. (Just use Table 12.1 and remember that $c = 0$.) The Cournot–Nash outputs lie on the 45° line, as do the output capacities (\bar{q}, \bar{q}) .

Depending on the level of exogenous capacity constraint parameter \bar{q} there are three types of solution to the price setting game. The three cases are:

- $\bar{q} \geq \alpha/\gamma$: each firm has enough capacity to supply the entire market at a price equal to the marginal production cost of zero;
- $(\alpha/3\gamma) < \bar{q} < \alpha/\gamma$: firm capacity is smaller than in case (a) but larger than the Cournot–Nash equilibrium output;
- $0 < \bar{q} \leq (\alpha/3\gamma)$: firm capacity is equal to or smaller than the Cournot–Nash output.

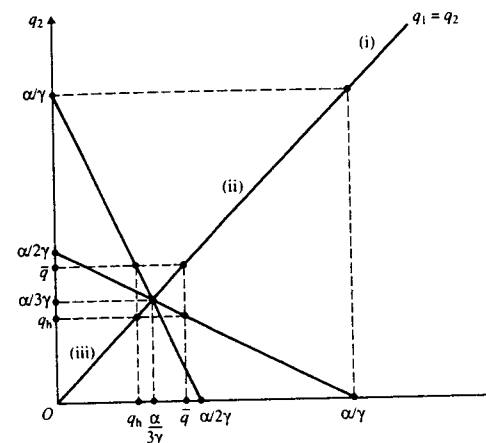


Fig. 12.4

In case (a) the Nash equilibrium solution of the capacity constrained price setting game is $p_1 = p_2 = 0$ which is the Bertrand solution when marginal production cost is zero. The capacity constraints have no effect on the solution because either firm can supply the entire demand at a price equal to marginal production cost.

The interesting cases are (b) and (c). Before we can analyse these we need to make two further assumptions. The first is *equal sharing*: if firms charge the same price they will each sell half the total quantity demanded at that price. The second is *efficient rationing*: if firm j sets a lower price than firm i and sells to its capacity \bar{q} , then firm i will face the residual demand curve $p_i = (\alpha - \gamma\bar{q}) - \gamma q_i$ shown by the segment $\hat{\alpha}(\alpha/\gamma)$ in Fig. 12.5. Rationing is

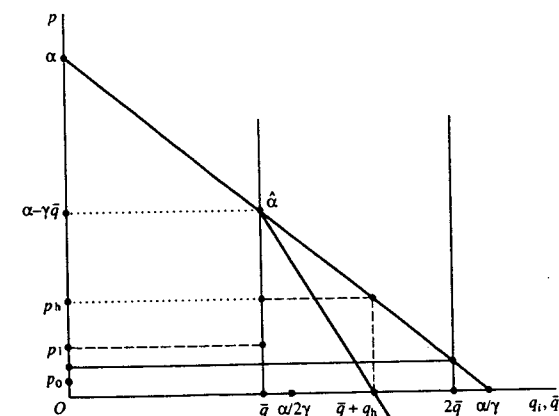


Fig. 12.5

efficient because it is as if j 's \bar{q} units of output were sold to those consumers who value them most highly, thus removing the upper segment $\alpha\hat{x}$ of the market demand curve and leaving firm i with the remainder. The efficient rationing assumption is rather strong since it is possible to think of other plausible means of allocating firm j 's output. For example, it could be allocated randomly to consumers, or to those who are first in line. As we note below, the form of rationing has a significant effect on the model's results.

The price p_0 at which demand is equal to the combined capacities of the firm is:

$$p_0 = \alpha - 2\gamma\bar{q}$$

Fig. 12.5 illustrates for case (b). In cases (b) and (c) neither firm will ever choose a price below p_0 . (If $\bar{q} > \alpha/2\gamma$ it would not be possible to set such a price.) If firm j has set $p_j \geq p_0$ then firm i can sell \bar{q} units at all prices $p_i \leq p_0$. Hence $p_i < p_0$ is not an optimal response to $p_j \geq p_0$ because raising p_i to p_0 would increase revenue. If firm j has set $p_j < p_0$ firm i will be able to sell \bar{q} units at $p_i = p_0$ and so $p_i < p_0$ is not an optimal response to $p_j < p_0$. Thus $p_i < p_0$ is a *dominated strategy* for firm i : it is not a best response to *any* strategy (p_j) of firm j .

The other significant feature of p_0 is that if firm j sets a price greater than p_0 the best response of firm i is not to set the same price. The equal sharing assumption implies that if $p_i = p_j > p_0$ firm i gets only half the market demand, which from the definition of p_0 is less than its capacity. By undercutting firm j slightly firm i can sell an output equal to its capacity and thereby earn more revenue than sharing the market. Thus a Nash equilibrium can never have $p_i > p_0$.

Now we consider case (b) and show that the best price response functions $p_i^*(p_j)$ are those graphed in Fig. 12.6. Since the response functions do not intersect there is no pair of prices which are best responses to each other: case (b) does not have a Nash equilibrium.

First note that a single firm faced with the demand function $\alpha - \gamma q$ would maximize revenue (and therefore profit since $c = 0$) by choosing either a quantity of $\alpha/2\gamma$, if this output was feasible, or an output equal to capacity if, as in Fig. 12.5, $\bar{q} < \alpha/2\gamma$. Consider

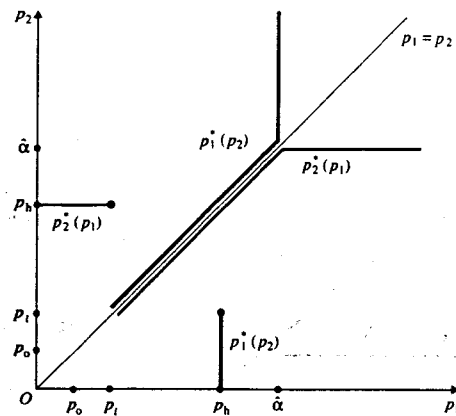


Fig. 12.6

the sub-case in which $\bar{q} < \alpha/2\gamma$ and denote the price at which demand equals the capacity of a single firm by

$$\hat{x} = \alpha - \gamma\bar{q}$$

(See Fig. 12.5.) Then \hat{x} is the best response by firm i to firm j setting any price greater than \hat{x} : firm i gets all the demand because it has the lower price and \hat{x} is its capacity constrained revenue maximizing price given $p_j > \hat{x}$. If firm j sets $p_j = \hat{x}$ firm i 's best response is to set a very slightly lower price $p_i = \hat{x} - \epsilon$. (If it sets $p_i > \hat{x}$ it sells nothing and we showed above that it does better undercutting rather than equalling $p_j = \hat{x} > p_0$.) Thus the best response function of firm 2 graphed in Fig. 12.6 as $p_2^*(p_1)$ is horizontal at \hat{x} for $p_1 > \hat{x}$ and lies ϵ below the 45° line at $p_1 = \hat{x}$. Similarly, firm 1's best response curve $p_1^*(p_2)$ is vertical at \hat{x} for $p_2 > \hat{x}$ and lies ϵ above the 45° line at $p_2 = \hat{x}$.

In the other sub-case where $\alpha/2\gamma \leq \bar{q}$ the best response curve for firm 2 is horizontal at the unconstrained revenue maximizing price $\frac{1}{2}\alpha$ for $p_1 > \frac{1}{2}\alpha$ and ϵ below the 45° line at $\frac{1}{2}\alpha$. The best response by firm i to $p_j \in [\hat{x}, \frac{1}{2}\alpha]$ is to set p_i just slightly less than p_j . (If it sets a higher price it gets no revenue and we have already established that setting the same price cannot be a best response.) Thus in both sub-cases the best response functions do not intersect at any price greater than or equal to the price \hat{x} at which demand is equal to the capacity of a single firm. Bertrand price competition rules out any equilibrium with prices of \hat{x} or greater.

Before we can consider best responses to prices below \hat{x} , we must define two other price levels p_h and p_ℓ which play a crucial role in the argument. Denote by p_h the price that firm i would set given that firm j sets a lower price and produces its capacity output. In these circumstances the efficient rationing assumption implies that firm i will face the residual demand curve $\hat{x}(\alpha/\gamma)$ and residual marginal revenue curve $\hat{x}(\bar{q} + q_h)$ in Fig. 12.5. It would set the price p_h at which the quantity sold was q_h . (q_h is also the best quantity response to firm j 's choice of \bar{q} . Thus in terms of Fig. 12.4, q_h is the point on firm i 's best response function for $q_j = \bar{q}$.) Define p_ℓ by

$$p_\ell \bar{q} = p_h q_h \quad [\text{B.28}]$$

The firm would be indifferent between selling its capacity output at p_ℓ and selling q_h at p_h . The reader should check that $p_h > p_\ell > p_0$ as in Fig. 12.5.

We can now show that firm i 's best response to $p_j \in [\hat{x}, p_h]$ is to undercut p_j slightly: $p_i = p_j - \epsilon$. If firm i responds to $p_j \in [\hat{x}, p_h]$ by setting $p_i^+ > p_j$ it must have $p_i^+ > p_h$. It faces the residual demand curve and sells $q_i^+ < \bar{q}$, which yields revenue $p_i^+ q_i^+ < p_h q_h$. (Remember p_h maximizes its revenue if it faces the residual demand curve.) If firm i instead sets $p_i^- = p_j - \epsilon > p_\ell$ it sells \bar{q} . For small enough ϵ this is better than setting the same price as firm j and getting half the market demand, which, since $p_i^- > p_0$, is less than $2\bar{q}$. Because $p_i^- > p_\ell$, [B.28] implies

$$p_i^- \bar{q} > p_\ell \bar{q} = p_h q_h > p_i^+ q_i^+$$

Hence firm i 's best response to $p_j \in [\hat{x}, p_h]$ is to undercut p_j slightly.

If firm j sets $p_j \in (p_h, p_\ell)$ firm i 's best response is again to undercut it slightly. Its best price greater than p_j is p_h . By setting $p_i = p_j - \epsilon > p_\ell$ firm i sells its capacity output, yielding greater revenue than a price of p_h :

$$(p_j - \epsilon)\bar{q} > p_\ell \bar{q} = p_h q_h$$

Thus over the range $[\hat{a}, p_l)$ firm 2's best response function lies ε below the 45° line and firm 1's is ε above it.

Firm i 's best response to $p_j \in [p_l, 0]$ is to set $p_i = p_h$. We know that matching firm j 's price $p_j > p_0$ is never optimal for firm i and setting a price ε less than p_j yields smaller revenue than choosing p_h since now

$$(p_j - \varepsilon)\bar{q} < p_l\bar{q} = p_h q_h$$

Over the range $[p_l, 0]$ firm 2's best response function is the horizontal line at $p_2 = p_h$ and firm 1's is the vertical line at $p_1 = p_h$.

Notice in Fig. 12.6 that because the best price response curves are discontinuous at p_l they never intersect and there is no pair of prices which are best responses to each other. Edgeworth viewed this market as a process taking place over successive time periods and argued that prices would cycle endlessly without reaching an equilibrium. No matter where the process started it would eventually lead to prices cycling over the range $[p_l, p_h]$. From p_h a process of competitive undercutting would drive price down to p_l , there would then be a jump back to p_h and the undercutting process would start again. However, here we view the market as a one shot game, with each firm reasoning through the 'process' in an attempt to formulate an optimal response to the other firm's price. We cannot then predict what decisions they will take because there is no Nash equilibrium.

This pessimistic result arises because we required the equilibrium to be a pair of prices which were best responses to each other. In game theoretic terms the players were assumed to be restricted to *pure strategies*: they had to choose a definite price (strategy) in response to a definite price (strategy) of the other player. There are many other examples in economics of games which do not possess Nash equilibria in pure strategies. The solution to this analytical difficulty is to introduce the concept of a *mixed strategy*. A mixed strategy is a probability distribution over pure strategies, specifying the probability with which each pure strategy will be chosen. (Note that the concept includes pure strategies as a special case: by setting the probabilities attached to all pure strategies except one equal to zero a player can choose one of the pure strategies for sure.) It can be shown that for the type of game we are analysing there must always exist a mixed strategy Nash equilibrium: a pair of probability distributions over pure strategies which are best responses to each other.

Before we show how to find the mixed strategy equilibrium in case (b) we shall analyse case (c) in which the exogenously given fixed capacities are at or below the Cournot-Nash equilibrium output levels. The same argument as in case (b) establishes that there cannot be any equilibrium in prices with price above p_h . Remember that p_h is the optimal price for firm i given that firm j will sell $q_j = \bar{q}$. The reader should check that in case (c) the capacity constraint binds at p_h so that $q_h = \bar{q}$ and

$$p_h = \hat{a} - \gamma\bar{q} = (\alpha - \gamma\bar{q}) - \gamma\bar{q} = \alpha - 2\gamma\bar{q} = p_0$$

Thus $p_l = p_0$ is the best response to any price set by firm j which results in j selling its capacity. But this means that $p_l = p_0$ is also the best response to $p_j = p_0$. Hence we have a Nash equilibrium in prices with $p_1 = p_2 = p_0$ and both firms selling their capacity outputs. Thus the Edgeworth non-existence of an equilibrium in pure strategies occurs only in case (b).

In terms of Fig. 12.4 the best unconstrained output response by firm 1 to firm 2 setting $q_2 = \bar{q}$ is at the intersection of the horizontal line at $q_2 = \bar{q}$ with its best output response

		Player 1 strategies	
		l	r
Player 2 strategies	u	5, 10	10, 2
	d	4, 8	2, 4

Fig. 12.7

curve. But this would violate firm 1's capacity constraint and firm 1 maximizes its profit, given $q_1 = \bar{q}$, making q_1 as large as possible: it moves along the horizontal line at $q_2 = \bar{q}$ until it reaches the capacity constraint. Thus when the firms' capacities are less than their Cournot-Nash equilibrium outputs the solution is at the intersection of the capacity constraints. Notice that if the capacity constraints were equal to the Cournot-Nash equilibrium outputs the Bertrand capacity constrained price setting game yields the same result as the unconstrained Cournot quantity setting game.

Mixed strategies: a brief exposition

The payoff matrix in Fig. 12.7 describes a simple one-shot game in which two players each have two pure strategies. They must choose simultaneously. The cells show the payoffs to the players resulting from the different possible combinations of pure strategies. The first number in each cell is the payoff to player 1 and the second the payoff to player 2. For example if player 1 chooses strategy r and player 2 chooses strategy u , player 1 gets 10 and player 2 gets 2. If we restrict the players to choosing one of their pure strategies there is no Nash equilibrium (NE). The pair (r, u) is not a NE: player 1's best response to u is l . But (r, u) is not an equilibrium since player 2's best response to r is d . Similarly (r, d) is not a NE since player 1's best response to d is l . Finally (l, d) is also not a NE since player 2's best response to l is u . Thus, if we imagined an Edgeworth-like process the players would cycle endlessly through the four strategy pairs without reaching a pair which are best responses to each other. There is no pure strategy NE in this case. (The reader should check that if the payoff to player 2 from the strategy pair (l, u) was changed to 6 there would be a NE at (l, d) .)

Now let us suppose that the players can choose the probability with which to play each of their pure strategies. For example, if player 1 decides to play l and r with equal probabilities, she could toss a fair coin. If she decided to play l with probability $2/13$ she could shuffle and cut a pack of cards and play l if the Ace or King turned up and r otherwise. Whatever the probabilities, we assume that a suitable randomizing device is available. Let us denote the probability that player 1 chooses l by x , so that she chooses r with probability $(1 - x)$. Similarly y is the probability that player 2 chooses u . A mixed strategy for player 1 is a choice of x which determines the probability with which she plays her l and r pure strategies. Clearly games in mixed strategies includes games in pure strategies as special cases where x and y are restricted to be 0 or 1.

Given the chosen probabilities the *expected payoffs* V^i to the players are

$$V^1(x, y) = 5xy + 4x(1 - y) + 10(1 - x)y + 2(1 - x)(1 - y)$$

$$V^2(x, y) = 10xy + 8x(1 - y) + 2(1 - x)y + 4(1 - x)(1 - y)$$

A *mixed strategy Nash equilibrium* is a pair of mixed strategies (x^*, y^*) with the property that $V^1(x^*, y^*) \geq V^1(x, y^*)$ for all $x \in [0, 1]$ and $V^2(x^*, y^*) \geq V^2(x^*, y)$ for all $y \in [0, 1]$. The mixed strategies must be best responses to each other.

The partial derivative of V^1 with respect to x is

$$V_x^1(x, y) = 5y + 4(1 - y) - 10y - (1 - y)2 = 2 - 7y$$

which is positive, zero or negative as y is less than, equal to or greater than $2/7$. Thus player 1's best response to $y < 2/7$ is to set $x = 1$, and to $y > 2/7$ is to set $x = 0$. If $y = 2/7$ player 1 would get the same expected payoff from choosing any $x \in [0, 1]$. Fig. 12.8 plots player 1's best mixed strategy response to y .

The partial derivative of V^2 with respect to y is

$$V_y^2(x, y) = 10x - 8x + 2(1 - x) - 4(1 - x) = 4x - 2$$

Player 2's best response to x less than $1/2$ is $y = 0$, to x greater than $1/2$ is $y = 1$. If $x = 1/2$ player 2 would be indifferent among all $y \in [0, 1]$. His best mixed strategy response to x is also shown in Fig. 12.8.

It is apparent from Fig. 12.8 that there is an equilibrium in mixed strategies for this one shot game at $(x^*, y^*) = (1/2, 2/7)$. This pair of mixed strategies are best responses to each other: if player 1 chooses to play her pure strategy l with probability $1/2$ then player 2 cannot do better than playing u with probability $2/7$. But, faced with player 2 choosing u with probability $2/7$, player 1 cannot do better than playing l with probability $1/2$.

It is possible to show that a Nash equilibrium in mixed strategies exists for *all* games in which the players have a finite number of pure strategies and for a wide class of games in which, as in the Edgeworth model, they have a continuum of strategies.

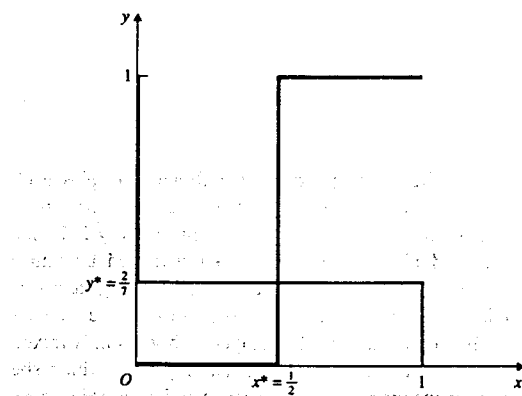


Fig. 12.8

The alert reader will notice that faced with $y = 2/7$ player 1 gets the same expected payoff whatever her choice of x . In order to predict that she will actually choose $x^* = 1/2$ we would have to extend our specification of the model beyond that shown in the payoff matrix. For example, we could assume that other things being equal players prefer to make choices which sustain an equilibrium. In some circumstances it may be appropriate to think of populations of players of each type. A proportion of type 1 players always choose l and the rest always choose r . Similarly a proportion of type 2 players always choose u and the rest always choose d . If we also specify some kind of evolutionary process which adjusts the proportions of both types of players when there is disequilibrium, the mixed strategy equilibrium has some plausible appeal as a stable state in the evolutionary process. Which story we use to justify predicting that the mixed strategy equilibrium strategies will actually be chosen depends on the context.

It follows from the fact that the expected payoff of a player is linear in probabilities that, even in more complex games, a mixed strategy equilibrium in which a player chooses a particular pure strategy with a probability which is neither zero nor 1 will always be characterized by the player getting the same expected payoff whatever probability she attaches to that pure strategy. We will now use this property to find the mixed strategy equilibrium of the Edgeworth model.

Mixed strategy equilibrium in Edgeworth's duopoly model

We saw earlier that for the case in which the exogenously fixed capacity output \bar{q} lies between α/γ (demand at zero price) and $\alpha/3\gamma$ (Cournot-Nash output q_i^c) Edgeworth's model has no Nash equilibrium price pair. Since in this model the firms' strategies are prices, this is equivalent to saying that the model has no pure strategy Nash equilibrium. We now show that it does have an equilibrium in mixed strategies, by actually calculating the probability distributions over prices that define this equilibrium. There is also a simple characterization of the equilibrium expected payoff to each firm.

Fig. 12.9 gives an alternative way of finding the interval $[p_l, p_h]$ which, as we will show, is the only set of prices chosen with positive probability in the mixed strategy equilibrium. The curve $R(p)$ is given by

$$R(p) = pq(p) = p(\alpha - p)/\gamma = (\alpha p - p^2)/\gamma$$

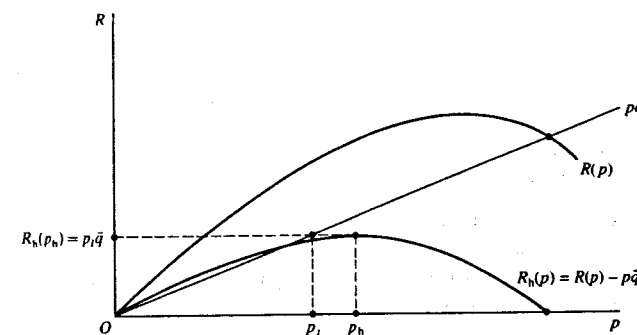


Fig. 12.9

where $q(p) = (\alpha - p)/\gamma$ is the demand function for total output $q = q_1 + q_2$. This is the revenue function a monopolist would face. Now if a firm is the higher priced firm, it faces the revenue function $R_h(p)$ defined by

$$R_h(p) = p(q(p) - \bar{q}) = [(\alpha - p^2)/\gamma] - p\bar{q} \quad [\text{B.29}]$$

since it receives the residual demand, after the lower-priced firm has sold its capacity output (recall the 'efficient rationing' assumption). We saw earlier that p_h maximizes $R_h(p)$ and so satisfies

$$(\alpha/\gamma) - 2p/\gamma - \bar{q} = 0 \Rightarrow p_h = (\alpha - \gamma\bar{q})/2 \quad [\text{B.30}]$$

We then have from [B.29] and [B.30] after some rearrangement

$$R_h(p_h) = \frac{1}{\gamma} \left(\frac{\alpha - \gamma\bar{q}}{2} \right)^2 \quad [\text{B.31}]$$

In the figure, p_h is shown as the price which maximizes $R_h(p)$. Then, p_l is found as the price which equates $p\bar{q}$ with $R_h(p_h)$, as in [B.28].

A mixed strategy for firm i is a distribution function $\psi_i(p)$ such that $\psi_i(p)$ is the probability that i chooses a price less than or equal to p . A mixed strategy equilibrium is a pair of distribution functions (ψ_1, ψ_2) which are best replies to each other. We will show that there is a mixed strategy equilibrium in which the distribution functions are identical and increasing in $p \in [p_l, p_h]$. Thus the probability density functions ψ'_i are positive over this interval.

If firm i chooses the mixed strategy $\psi_i(p)$, then ignoring some technical issues (for which see Kreps and Scheinkman (1983)), when j sets price p it will be the lower price firm with probability $1 - \psi_i(p)$ and the higher price firm with probability $\psi_i(p)$. Thus j 's expected revenue if it sets price p is

$$ER_j(p) = [1 - \psi_i(p)]p\bar{q} + \psi_i(p)R_h(p) \quad [\text{B.32}]$$

Now we are looking for an equilibrium in which all prices in the interval $[p_l, p_h]$ may be chosen by firm i . But this requires that all prices in the interval yield j the same expected revenue when firm i chooses the distribution function $\psi_i(p)$. If there was a price which yielded a unique global maximum over this interval firm j would choose it for sure. Any price which yields a smaller expected revenue than some other price in the interval would never be chosen. Thus all prices p must yield the same expected revenue to j if the requirement that all prices in the interval have a positive probability of being chosen ($\psi'_j(p) > 0$) is to be satisfied.

By choosing p_l , j can get expected revenue of

$$ER_j(p_l) = [1 - \psi_i(p_l)]p_l\bar{q} + \psi_i(p_l)R_h(p_l) = p_l\bar{q}$$

since $\psi_i(p_l) = 0$. Thus firm i 's distribution function ψ_i must be such that for all $p \in [p_l, p_h]$ firm j gets an expected revenue of $p_l\bar{q}$:

$$ER_j(p) = p_l\bar{q} = [1 - \psi_i(p)]p\bar{q} + \psi_i(p)R_h(p) \quad p \in [p_l, p_h] \quad [\text{B.33}]$$

Solving this for $\psi_i(p)$ gives

$$\psi_i(p) = (p\bar{q} - p_l\bar{q})/[p\bar{q} - R_h(p)] \quad p \in [p_l, p_h] \quad [\text{B.34}]$$

To see that this indeed gives a strictly increasing distribution function, note that

- (a) if $p = p_l$, then $\psi_i(p_l) = 0$
- (b) if $p = p_h$, then $\psi_i(p_h) = 1$ (since $R_h(p_h) = p_l\bar{q}$)
- (c) for $p_l < p < p_h$, $\psi_i(p) > 0$ (since then $p\bar{q} > R_h(p)$, as can be confirmed from Fig. 12.9)
- (d) $\psi'_i > 0$ if and only if (differentiating through [B.34] and cancelling terms),

$$[p_l\bar{q} - R_h(p)] + R'_h \cdot [p - p_l] > 0 \quad [\text{B.35}]$$

which Fig. 12.9 readily confirms will hold.

If firm i chose the strictly increasing distribution function given by [B.34] then firm j would be willing to randomize over its price and would, in particular, be willing to choose the same distribution function as firm i . Thus we have a symmetric mixed strategy Nash equilibrium in which the distribution function [B.34] is a best reply to itself.

Not all economists find the idea that firms choose prices by use of a randomizing device an appealing one. This result does have one interesting implication, however. It is often argued, particularly in anti-trust cases, that the observation of identical prices set by oligopolistic firms, and the tendency of those prices to change virtually simultaneously by identical amounts, is not evidence of collusion *per se*, since it would also be observed in a competitive market. In a market for which the mixed strategy Edgeworth equilibrium holds, however, this would *never* be observed: prices are identical with zero probability, and would change stochastically (if we think of repeated plays of the same game). Firms are however behaving non-collusively, since they are choosing prices independently of each other. If a market fitted this case *and* we observed identical prices, then we could conclude that the firms were behaving collusively.

Prices vs. quantities and the Kreps-Scheinkman model

The foregoing survey of duopoly models shows that whether firms are assumed to choose prices or quantities is of considerable importance. Bertrand thought it obvious that firms should be regarded as price-setters and indeed thought that Cournot's specification in terms of quantity choice was simply an analytical mistake. One prevalent modern view seems to be that no such judgement need be made, and that it is an empirical matter to decide which type of model better fits a particular market. In the market being analysed, do firms set outputs and then allow price to adjust to whatever level allows them to be sold, or do they set prices and then produce to meet whatever demands arise? The answer determines which type of model to use. A second strand in the literature considers the question of the *endogenous* choice of strategy variable. An important paper somewhat in this spirit is that of Kreps and Scheinkman (1983). This paper could be thought of as an extension of the Edgeworth model just considered, in that it allows the firms' capacity levels to be *endogenously chosen*. In doing so the paper makes an interesting reconciliation between the Cournot and Bertrand models. It is shown that if firms first choose capacity outputs and then, with these capacities fixed, play a price-setting game of the kind just

analysed as the Edgeworth model, then the equilibrium of the model takes the following form. The equilibrium capacities at the first stage correspond to the Cournot–Nash equilibrium outputs for the market, $\bar{q} = q_i^c$. The second stage price game consists of the firms setting their prices at the level at which demand equals the sum of Cournot–Nash outputs. Thus we have the Cournot–Nash equilibrium in a market in which firms set prices, subject to precommitted capacity levels.

Unfortunately, this striking result is not robust to variations in the assumption on the form of rationing by the lower priced firm in this model. As Davidson and Deneckere (1985) show, if some other assumption is made than that referred to above as the ‘efficient rationing assumption’, which defines the residual demand function for the higher-priced firm as $p_i = (\alpha - \gamma\bar{q}) - \gamma q_i$, then the Kreps–Scheinkman result no longer holds. The choice between prices or quantities as strategy variables matters.

Exercise 12B

- Given the inverse demand functions in [B.2], show that the parameters of the demand function in [B.5] are:

$$a_i = \frac{\beta_j x_i - \gamma x_j}{\beta_1 \beta_2 - \gamma^2}; \quad b_i = \frac{\beta_j}{\beta_1 \beta_2 - \gamma^2}; \quad \phi = \frac{\gamma}{\beta_1 \beta_2 - \gamma^2} \quad i, j = 1, 2 \quad i \neq j$$

for the case where outputs are not homogeneous. Why is it necessary to assume that $\beta_j x_i > \gamma x_j$ and $\beta_1 \beta_2 > \gamma^2$?

- Use Table 12.1 to prove [B.16] in the case of non-homogeneous outputs.
- Explain why, in Table 12.1, the individual outputs q_i^m and \hat{q}_i are indeterminate, when firms’ outputs are homogeneous.
- Show that q_i^R and q_i^m as defined in the text satisfy $q_i^R > q_i^m$.
- Confirm the construction of Fig. 12.4, and hence the unique Bertrand equilibrium (p_1^R, p_2^R) , by showing that $\bar{A}_i > 0$, $1 > \bar{B}_i > 0$, $i = 1, 2$. (Hint: use the facts and assumptions in Question 1.)
- Show that (p_1^R, p_2^R) imply positive profits, but a more competitive outcome than (q_1^c, q_2^c) , in the model of this section with non-homogeneous outputs.
- Show that the profit functions in [B.4] are strictly quasi-concave. (Hint: use the expression for dq_2/dq_1 given in the discussion of the Stackelberg model.)
- Show that in case (a) of the Edgeworth model the Nash equilibrium of the price setting game has each firm setting a price of zero. Draw the best price response functions and confirm they intersect at the origin.

C. Oligopoly as a repeated game

We retain the basic duopoly model set out at the beginning of the previous section, but now the firms choose prices or outputs in each of a sequence of time periods. The game played in each period is the *constituent game*, and it is common knowledge to the firms

that they are engaged in a sequence of repetitions of this game. They formulate strategies for the repeated game, not just for the one-period constituent game in isolation from any other period.

In such a context it is possible to rationalize collusive behaviour in the absence of binding (legally enforceable) agreements. If the firms agree, explicitly or tacitly, to collude in one period, and if one firm then deviates from that agreement, the other can punish it by instigating a price war (output expansion) or carrying out some other retaliatory action in the next period. The threat of anticipated future punishment may make it rational for each firm to adhere to the agreement. The firms can make an agreement in the belief that it will be sustained by self interest: it is *self-enforcing*.

This simple and appealing idea must however be subjected to further analysis. First, the gains from deviation will be realized ‘now’, while the losses from punishment will occur in the future. Will it always be the case that sufficiently large future losses can be threatened to offset the gains from immediate deviation? This depends on the mechanism by which punishment is inflicted, the rate at which firms discount future profits, and the length of time for which the deviant can gain from breaking the agreement before punishment begins. A second issue is the credibility of the threat of punishment. Typically, inflicting punishment through the market – for example by a price war – hurts the punisher as well as the deviant. The threat of punishment will be an effective deterrent only if potential deviants believe that it will actually be carried out. These questions are the central concern in the models we consider in this section and the next.

We first need to consider whether the constituent game is repeated a finite number of times or infinitely often. If there is a known last period of the game backward induction shows that the above intuitive argument for collusion may break down. The equilibrium repeated game strategy may then simply consist of repeated plays of the one-shot Nash equilibrium. For example, consider the differentiated product model of the previous section in which firms choose prices. In the last period of the repeated game, the usual argument establishes that firms choose the (unique) Bertrand–Nash equilibrium prices. There is no next period in which to punish deviation and collusive prices are not a Nash equilibrium of this one-shot game.

In the next-to-last period, the firms could agree to collude, but it is not possible to support this with credible threats of punishment in the last period. For the *sub-game of the repeated game* consisting of the last-period one-shot game, the only Nash equilibrium is the Bertrand equilibrium. In the next to last period both firms realize that this is the case. So the threat of punishment by setting a price in the last period other than the Nash equilibrium price is not credible and cannot sustain collusion in the next to last period. But if collusion cannot be sustained the Nash equilibrium in the next to last period is the Bertrand price equilibrium. This in turn implies that the threat of setting non-Bertrand prices in the next to last or the last period cannot sustain collusion in the second to last period game and Bertrand equilibrium prices are the only Nash equilibrium in the second to last period. The argument extends, period by period, right back to the first. Thus the only credible Nash equilibrium of the finitely repeated game has the firms choosing the Nash equilibrium of the one-shot game in every period.

If the game is repeated forever, there is no last period in which to start the backward induction process. The repeated game will look exactly the same from whatever point in time it is considered. In this case, as we shall see, collusion can be rationalized as a Nash equilibrium of the repeated game.

In the remainder of this chapter we assume an infinitely repeated market game. Although individuals have finite lives, firms as institutions have potentially infinite lives, and the individuals within them who take decisions realise this. Moreover, it can be shown that if a repeated game has a finite, but uncertain, number of periods, then collusion may be sustainable. If, in a given period, there is some probability that there will be a next period, then there will be an expected value of loss from punishment which might sustain collusion. In effect, the probability of a future can be incorporated into the time discount factor.³

We take the model of the previous section in both the differentiated and homogeneous product cases. The time period is denoted $t = 0, 1, \dots, \infty$, and choices of prices or outputs are made in each period. We assume that the firms face the same interest rate $r > 0$, and wish to maximize the present value of profit.

$$V_i = \sum_{t=0}^{\infty} \delta^t \pi_i^t \quad [\text{C.1}]$$

where $\delta \equiv (1 + r)^{-1}$ is each firm's discount factor and π_i^t is i 's profit in period t . In this discounting formula we assume that profits accrue at the beginning of each period, which is also when decisions are taken. Before considering firms' strategies for the repeated game, we need to extend our previous analysis of the constituent game.

It will be useful for illustrative purposes to have numerical versions of the models defined earlier in [B.1]–[B.5]. We assume:

$$\alpha_1 = \alpha_2 = 10; \quad c_1 = c_2 = 1$$

$$\text{Differentiated products case: } \beta_1 = \beta_2 = 1; \quad \gamma = 0.5$$

$$\text{Homogeneous products case: } \gamma = 1$$

Thus the firms have identical cost and demand functions. We are interested in the Nash equilibria of the one-shot game for both quantity and price choices, as well as in the joint profit maximizing solution. Table 12.2 gives the values of prices, outputs and profits at these various solutions for this numerical example. The numbers of course confirm the earlier algebraic results. Note that in the homogeneous products case, since individual outputs are indeterminate in the Bertrand–Nash and joint-profit maximization (monopoly) cases, only total outputs are given. In the former case individual profits are necessarily zero since $p^B = c$, while in the latter case individual profits are *a priori* indeterminate and the total profit is given.

If we assume that the firms collude, the joint profit maximizing allocation is a natural one to focus upon, because it gives the maximum gain they can make from their cooperation. However, we should not think of it as the only possible collusive outcome,

Table 12.2

	Differentiated products	Homogeneous products
Cournot–Nash	$q_i^C = 3.6$; $p_i^C = 4.6$; $\pi_i^C = 12.96$	$q_i^C = 3$; $p^C = 4$; $\pi_i^C = 9$
Bertrand–Nash	$q_i^B = 4$; $p_i^B = 4$; $\pi_i^B = 12$	$q^B = 9$; $p^B = 1$; $\pi_i^B = 0$
Joint-profit maximization	$q_i^M = 3$; $p_i^M = 5.5$; $\pi_i^M = 13.5$	$q^M = 4.5$; $p^M = 5.5$; $\pi^M = 20.25$

in the absence of a well-defined model that would predict it to be so. Moreover, it is not difficult to construct models in which the Cournot–Nash equilibrium output yields a higher profit for one of the firms than it would obtain from the output it would produce at the joint profit maximizing equilibrium (see Question 1, Exercise 12C). Such a firm would not then agree to move from the former to the latter unless *side-payments*, or lump-sum redistributions of profit between the two firms, are feasible. If they are, the firms maximize their gains from collusion by producing outputs q_i^M in the differentiated products case and q^M in the homogeneous products case to generate a total profit π^M , and then making whatever side-payment from one to the other is necessary to achieve agreement. The relationship between their actual profits π_i^L after sidepayments is

$$\pi_2^L = \pi^M - \pi_1^L \quad [\text{C.2}]$$

Suppose, however, that side-payments are not feasible. For example, in many countries collusion is illegal, and lump-sum side-payments would be strong evidence of collusion in a case against the firms. An alternative way to transfer profit is to vary outputs away from the levels q_i^M , with each firm retaining the profit it makes from sale of its own output. Two questions then arise: how should the firms do this? And, is it costly to them in the sense that total joint profits are lower than in the case where side-payments are feasible?

In the homogeneous product case, the answers to both questions are immediate. If the firms keep total output at the level q^M , so that price remains at p^M , then redistributing profit by output variation is equivalent to lump-sum redistribution. Since firms' marginal costs are constant and identical at c total profit does not depend on how the firms allocate a given total output between themselves:

$$\pi_i = (p^M - c)q_i \quad \text{and} \quad q_1 + q_2 = q^M \Rightarrow \pi_i = (p^M - c)q_i = \pi^M \quad [\text{C.3}]$$

This would not be true if the firms had non-constant marginal costs (whether or not they are identical), or constant but unequal marginal costs. In these cases, reallocation of outputs away from the joint profit maximum, increasing one firm's output and profit and reducing the other's, also reduces total profit because it violates the condition that total output be produced at minimum cost, a necessary condition for joint-profit maximization. For example, if the firms have unequal constant marginal costs then under joint profit maximization the lower cost firm should produce the entire output. (Question 2, Exercise 12C asks you to examine these cases further.)

In the differentiated products model the firms will do the best they can by reallocating outputs and profits so that for any given profit level of firm i , firm j 's profit is maximized. Formally, they solve the problem

$$\max_{q_1, q_2} \pi_2(q_1, q_2) \quad \text{s.t. } \pi_2(q_1, q_2) \geq \pi_1^0 \quad [\text{C.4}]$$

where π_1^0 varies over the interval $[0, \hat{\pi}_1]$, with $\hat{\pi}_1$ the level at which the value of π_2 obtained at the solution to [C.4] is zero. This restriction is imposed because neither firm would accept a negative profit if zero output and profit are always an option. The properties of the π_i functions discussed in section B ensure that solution values (q_1^* , q_2^*) for this problem exist and are unique for each π_1^0 . Since the solution is a function of the value of this constraint constant π_1^0 , the maximized value of firm 2's profit, π_2^* , is also a function of π_1^0 .

That is

$$\pi_2^* = \pi_2(q_1^*(\pi_1^0), q_2^*(\pi_1^0)) \equiv P(\pi_1^0) \quad [\text{C.5}]$$

For $\pi_1 \in [0, \hat{\pi}_1]$, the set of profit pairs $(\pi_1, P(\pi_1))$ then defines the market's *profit frontier*, giving the maximum level of one firm's profit for each level of the other's over a particular range. Since this frontier will play an important role in what follows we now examine it more closely.

The Lagrange function for the problem in [C.4] is

$$L(q_1, q_2, \lambda) = \pi_2(q_1, q_2) + \lambda[\pi_1(q_1, q_2) - \pi_1^0] \quad [\text{C.6}]$$

and the first-order conditions are

$$\pi_{2k}(q_1^*, q_2^*) + \lambda^* \pi_{1k}(q_1^*, q_2^*) = 0 \quad k = 1, 2 \quad [\text{C.7}]$$

$$\pi_1(q_1^*, q_2^*) = \pi_1^0 \quad [\text{C.8}]$$

where $\pi_{ik} = \partial \pi_i / \partial q_k$. Consider first the interpretation of λ^* . From the Envelope Theorem (Chapter 2, section J) we know that

$$\frac{\partial \pi_2^*}{\partial \pi_1^0} = \frac{\partial L}{\partial \pi_1^0} = -\lambda^* = P'(\pi_1^0) \quad [\text{C.9}]$$

so that $-\lambda^*$ is the slope of the profit frontier. Note that in [C.6], setting $\lambda = 1$ gives us the problem of maximizing the firms' joint profits, considered earlier in [B.14] (since addition of a constant π_1^0 to this problem makes no difference to its solution). Thus we can guess that the joint-profit maximizing profit pair, (π_1^m, π_2^m) is the point on the profit frontier at which its slope is -1 . This can be confirmed by setting $\lambda^* = 1$ in conditions [C.7] and noting that they are then identical to [B.15] and so will give the same output pair. Thus for $\lambda = 1$, $\pi_1^0 = \pi_1^m$.

Next, taking conditions [C.7] and eliminating λ^* gives the usual kind of tangency condition

$$\frac{\pi_{21}(q_1^*, q_2^*)}{\pi_{22}(q_1^*, q_2^*)} = \frac{\pi_{11}(q_1^*, q_2^*)}{\pi_{12}(q_1^*, q_2^*)} \quad [\text{C.10}]$$

since $-\pi_{11}/\pi_{12} = dq_2/dq_1$ is the slope of firm i 's profit contour. This condition defines the set of pairs (q_1^*, q_2^*) at the points of tangency of the firms' profit contours, as Fig. 12.10(a) illustrates. This figure is drawn for the numerical values of the parameters given earlier, so that

$$p_i = 10 - q_i - 0.5q_j \quad C_i = q_i \quad i = 1, 2 \quad i \neq j$$

are the underlying demand and cost functions.

In Fig. 12.10(b) the profit frontier is derived⁴ from the locus of tangency points in Fig. 12.10(a). The tangent line L to the frontier at the joint profit maximizing point $\pi^m = (\pi_1^m, \pi_2^m)$ has slope -1 and corresponds to $\lambda^* = 1$. The symmetry and concavity of both curves in Fig. 12.10 are due to the specific example chosen. There would be asymmetry if the firms had different profit functions, while the profit frontier in Fig. 12.10(b) can be made non-concave by choosing other, quite reasonable, functional forms for the demand

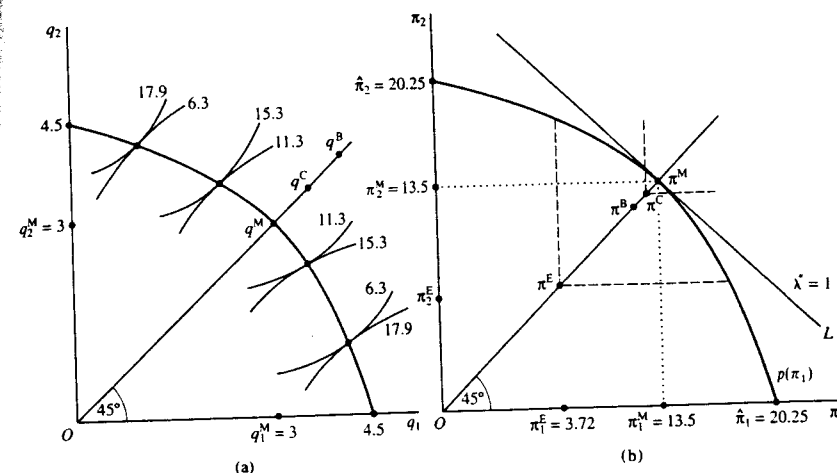


Fig. 12.10

or cost functions. The proof that for the functional forms used in this chapter the profit frontier is always concave is set as an exercise (see Question 3, Exercise 12C).

For comparison, the one-shot Nash equilibria in Table 12.3 are also shown in Fig. 12.10. Because of the symmetry in the example, the output and profit pairs lie on the 45° lines in the respective figures. The non-collusive equilibria lie inside the profit frontier and so are dominated by the allocations lying on the profit frontier north-east of them, as Fig. 12.11 illustrates in more detail. The Cournot-Nash equilibrium is dominated by all points on the arc cc' , and the Bertrand-Nash by all points on the arc bb' . The figure illustrates the firms' *incentive to collude*.

We can now answer the two questions put earlier. If the firms wish to achieve a profit allocation other than π^m , and if side-payments are possible, they can move along the line L , since this has the equation $\pi_2 = \pi^m - \pi_1$ (its slope is -1). If side payments are not possible then profit reallocation is costly, since the best the firms can do is to move along the curve $P(\pi_1)$, which must result in lower *total* profit than π^m .

Since in this example the firms have identical constant costs, the source of the loss in aggregate profit is the movement of outputs away from the values which equalize marginal revenues of the two outputs, where these marginal revenues take account of the effect of one output on the demand for the other. (Use condition [B.15] to show that this is a necessary condition for joint profit maximization.) The advantage of collusion is that it can internalize the 'external effect' that each firm's output has on the revenue of the other. Even if internalization is not complete (as it is under joint profit maximization), Fig. 12.10 shows that it can still achieve profit pairs which give higher profits to both firms than in the non-collusive equilibria.

We have established an incentive to collude, but will collusion be sustainable? We now consider some models which explore this question.

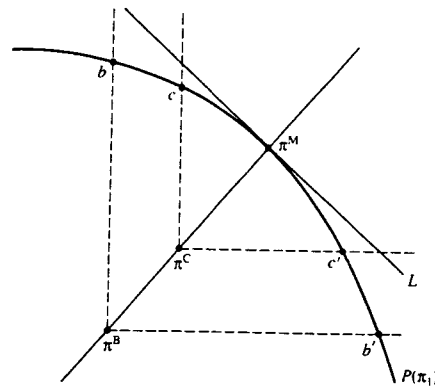


Fig. 12.11

Punishment by Cournot-Nash competition (J. W. Friedman)

Suppose that the firms agree to produce an output pair (q_1^*, q_2^*) that puts them at a point on the profit frontier $p(\pi_1)$ somewhere on the arc cc' in Fig. 12.11. Assume further that if instead they behaved non-collusively, then in each period of the repeated game they would be at the Cournot-Nash equilibrium, i.e. they are quantity-setters. To sustain their collusive agreement, the firms also agree on the following *trigger strategies*: for any $t = 0, 1, \dots$, if firm i produces q_i^* in period t , then firm j will produce q_j^* in period $t + 1$; however, if i reneges by producing $q_i^R \neq q_i^*$ in t , then j will produce its Cournot-Nash equilibrium output q_j^C in period $t + 1$ and *every succeeding period*. A deviation by one firm triggers a permanent switch by the other to its Cournot-Nash equilibrium output.

Suppose that at $t = 0$ firm i considers reneging on the agreement. Since it expects j to produce q_j^* , its best reneging output is $q_i^R = A_i - B_i q_j^*$, i.e. its best response to q_j^* . Let $\pi_i^R = \pi_i(q_i^R, q_j^*)$. Then its immediate gain is $\pi_i^R - \pi_i^*$. We know this to be positive because q_i^* is not the best response to q_j^* . However, under the trigger strategy, it will then be faced with q_j^C in every future period. Its best response to this is q_i^C , yielding profit π_i^C . Thus relative to the case in which it does not renege at $t = 0$, i will lose an infinite profit stream of $(\pi_i^* - \pi_i^C)$ with a present value of $(\pi_i^* - \pi_i^C)/r$, and so it will *not* pay i to renege at $t = 0$ if

$$\pi_i^R - \pi_i^* \leq (\pi_i^* - \pi_i^C)/r \quad [\text{C.11}]$$

or equivalently if

$$r \leq (\pi_i^R - \pi_i^C)/(\pi_i^* - \pi_i^C) \quad [\text{C.12}]$$

Since the repeated game is the same regardless of the t at which it begins, if [C.11] is satisfied at one t it is satisfied at all and so the trigger strategies will support the outputs (q_1^*, q_2^*) forever.

[C.11] says that i will not renege if its immediate profit gain is outweighed by the present value of future losses of profit. [C.12] expresses this condition in terms of an upper bound on the interest rate. Given the demand and cost parameters that determine the relations

among π_i^C , π_i^R and π_i^* , collusion will be sustainable provided the firms do not discount the future 'too heavily', thus weakening the force of the future punishment.

Since $r > 0$, [C.12] requires that $\pi_i^* > \pi_i^C$. Thus, in terms of Fig. 12.10, any profit pair along cc' (or indeed in the convex set defined by $\pi^c cc'$), such that this inequality is satisfied, is sustainable at *some* interest rate. For example consider the joint profit maximizing point $\pi^m = (13.5, 13.5)$. Firm j 's best response output to $q_i^m = 3$ is $q_j^R = 3.75$, yielding a profit $\pi_j^R = 14.06$. Thus applying [C.12] we have

$$r \leq (13.5 - 12.96)/(14.06 - 13.5) = 0.96 \quad [\text{C.13}]$$

Collusion at the joint profit maximum could be sustained in this market at any interest rate below 96 per cent per period.

Returning to the general case, if [C.12] is satisfied then the trigger strategies sustaining collusion represent a Nash equilibrium of the repeated game. If i believes j will play its trigger strategy, then i 's best response is to play its trigger strategy. Thus, the *equilibrium output path* will be (q_1^*, q_2^*) in every period. Are the threats underlying these trigger strategies – of playing the one-shot Nash equilibrium forever following a deviation – credible? In the next section we introduce two criteria for the credibility of threats and consider whether the trigger strategies satisfy them.

Exercise 12C

- Two firms produce homogeneous outputs with cost functions

$$C_1 = q_1^2 \quad C_2 = 2q_2^2$$

and the inverse market demand function

$$p = 100 - (q_1 + q_2)$$

Show that at the Cournot-Nash equilibrium firm 2 makes higher profit than at the joint-profit maximizing equilibrium. Explain why this is so.

- In the example of Question 1, derive the profit frontier, and explain why total profits fall as the firms redistribute profit between themselves by redistributing output. Then go through the same exercise replacing the cost functions of Question 1 by

$$C_1 = q_1 \quad C_2 = 2q_2$$

- Prove that for the functional forms assumed in this chapter the profit frontier is concave. Construct the profit frontier for the homogeneous output example of this section.

D. Credible threats*

Sub-game perfect equilibrium

We introduce the first of the criteria for equilibria of games involving threats with the simple example shown in Fig. 12.12. Player 1 moves first, and chooses one of two possible moves, u_1 or d_1 . Player 2 then moves, and chooses u_2 or d_2 in full knowledge of the choice 1 has

made. The numbered circles represent *decision nodes* for each player. The payoffs resulting from the four possible sequences of moves are as shown, with 1's payoff first. Part (a) of the figure shows the *extensive form*, which preserves the sequential aspect of the moves in the game, while (b) shows the *normal form* of the game, giving the matrix of payoffs for the possible combinations of strategies.

A strategy is a complete specification of the moves a player may make in the entire course of the game. For player 1, the strategies coincide with the moves: she chooses either u_1 or d_1 . Thus the payoff matrix in (b) of the figure has only two columns, one for each of 1's possible strategies. Player 2's strategies are a little more complex: a strategy must specify in advance exactly how he will move in each possible contingency. Thus he has four possible strategies

- s_1 : if u_1 then u_2 ; if d_1 then u_2 .
- s_2 : if u_1 then u_2 ; if d_1 then d_2 .
- s_3 : if u_1 then d_2 ; if d_1 then u_2 .
- s_4 : if u_1 then d_2 ; if d_1 then d_2 .

In other words a strategy for 2 must give instructions on how to play the game at *each possible* decision node in Fig. 12.12(a). The reader should confirm that the payoffs shown for each strategy pair in Fig. 12.12(b) are the end points of the corresponding paths through the game tree in Fig. 12.12(a).

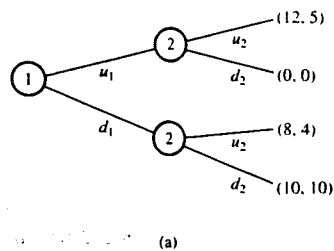
Referring to Fig. 12.12(b) we can quickly show that there are three Nash equilibrium strategy pairs, namely

$$(u_1, s_1), (u_1, s_2), (d_1, s_4)$$

To see this, note that

- s_1 or s_2 is 2's best response to u_1 , and u_1 is 1's best response to s_1 or s_2 ;
- s_4 is 2's best response to d_1 and d_1 is 1's best response to s_4 .

The reader should show that this 'mutual best response' property is possessed by no other strategy pair (in particular, (d_1, s_2)). The following argument suggests that of the three Nash equilibrium strategies only (u_1, s_2) is a reasonable prediction of the outcome. Consider the strategy pair (d_1, s_4) , and refer to Fig. 12.12(a). We could interpret this strategy as



	u_1	d_1
s_1	12, 5	8, 4
s_2	12, 5	10, 10
s_3	0, 0	8, 4
s_4	0, 0	10, 10

Fig. 12.12

embodying a threat. Player 2 naturally prefers the outcome (10, 10) and is threatening 1 that if she plays u_1 rather than d_1 , he will play d_2 , thus giving her 0, when she could have 10, by playing d_1 .

Is this a credible threat? If 1 chose u_1 , then 2 will have to do the best he can given that fact. But clearly in that case, 2 does better by choosing u_2 than d_2 . So, 2's threat is not credible because he has no incentive to carry it out when he is placed in the position to do so.

We also reject (u_1, s_1) as a reasonable strategy pair, since it requires 2 to choose u_2 if 1 chooses d_1 , when d_2 is clearly better for 2. In this case, (u_1, s_1) would result in the same outcome as (u_1, s_2) , since it prescribes the same choice for 2 in the event that 1 chooses u_1 , which she does in this strategy pair. Nevertheless, if we require that a strategy must specify rational behaviour in any contingency, whether or not that would actually arise in a play of the game, then (u_1, s_1) should also be ruled out.

On the other hand, (u_1, s_2) has 2 behaving rationally in all contingencies and it is the only Nash equilibrium strategy pair which would not be rejected as unreasonable.

The concept of *sub-game perfect equilibrium*, introduced by R. Selten, generalizes the idea of 'reasonableness' underlying this example. Its effect is often to narrow down the equilibria of a game to a subset of the Nash equilibria, and for that reason it is referred to as a 'refinement' of Nash equilibrium. In a game with a sequence of moves, take a decision node at any point in the game (including the first node) and identify the sequence of moves after that point as a game in itself: a *sub-game* of the original game. For example in Fig. 12.12 there are three sub-games: the games beginning at each of 2's decision nodes (which consist only of one move by 2), and the original game itself. This simple game does not have a very rich structure of sub-games. On the other hand, think of the infinitely repeated duopoly game beginning at $t = 0$. The players will make choices again at $t = 1$, $t = 2, \dots$, and so there is an infinity of sub-games of the game beginning at $t = 0$, namely those beginning at $t = 1, t = 2, \dots$ (as well as the game itself), and moreover, each of these sub-games is identical to the original game.

The formal definition of sub-game perfect equilibrium is that the players' strategies for a game must induce a Nash equilibrium in *every sub-game* of that game. Since a game is a sub-game of itself, this implies that a sub-game perfect equilibrium must be a Nash equilibrium. However some Nash equilibria of a game may not be sub-game perfect, because the strategies underlying them do not prescribe Nash equilibrium behaviour in all sub-games. The requirement of Nash equilibrium behaviour in sub-games ensures the credibility of threats: a threat which corresponds to playing a Nash equilibrium strategy for a sub-game beginning at a particular node *within* the game is credible, while an action that is not consistent with Nash equilibrium in the sub-game will not be a credible threat at the beginning of the game. Thus in the example of Fig. 12.12 the threat to choose d_2 following u_1 was not credible because it was not 2's best action for the sub-game beginning at the upper node 2.

Credibility of trigger strategies

Collusion supported by trigger strategies embodying punishment by Cournot-Nash competition is a sub-game perfect equilibrium and in that sense the threat of punishment is credible. Thus suppose that firm i observed that j has reneged at period t . Its trigger

strategy prescribes that it should produce q_i^t in every period from $t + 1$ onward. Firm j 's best response to q_i^t is to produce its Cournot–Nash output q_j^t . But choice of the output pair (q_i^t, q_j^t) in every period is a Nash equilibrium of this sub-game – the outputs are mutually best responses – and so the punishment strategies satisfy the requirement of sub-game perfection.

Nonetheless we may still doubt the reasonableness of these trigger strategies. Eternal punishment seems excessively grim, and we may feel that a punishment which ‘fits the crime’ would be a more plausible outcome. Punishment also hurts the punisher, in the sense that the Cournot–Nash output is less profitable for firm i than some collusive outputs. We might then expect firm j to propose to i that it should ‘forgive and forget’ and revert to cooperation. But if *ex ante* such renegotiation of the trigger strategies were anticipated to be successful, the credibility of the threat of punishment would be undermined. An alternative criterion of credibility of punishment strategies has therefore been proposed by J. Farrell and E. Maskin, that of *renegotiation-proofness*. We consider it at the end of this section.

Another difficulty with punishment by Cournot–Nash competition is that it may not be very severe if the Cournot–Nash profit is close to the collusive profit frontier as in Fig. 12.10(b). Since $r > 0$, condition [C.12] requires that $\pi_i > \pi_i^c$ if π_i is to be sustained. In Fig. 12.10(b), this means that only the relatively small set of points on and below the profit frontier and within the dashed lines drawn from π^c are sustainable at some $r > 0$. The set of r -values for which any given collusive allocation is sustainable would be widened if a more severe punishment than π^c could be inflicted. We now turn to punishment strategies which dramatically expand the possibilities of collusion.

The Folk Theorem

It is part of the conventional wisdom of game theory that threats of *minimax punishments* can sustain any *individually rational* collusive allocation as a Nash equilibrium of an infinitely repeated game. Since it is not possible to assign authorship of the result, it is known as the Folk Theorem.

A *minimax punishment* is the worst one firm can do to the other given that the firm being punished is making its best response to the action of the punisher. Supposing for definiteness that firm 1 is punishing firm 2, the output pair (q_1^t, q_2^t) has firm 1 minimaxing firm 2 if it solves

$$\min_{q_1} \max_{q_2} \pi_2(q_1, q_2) \quad [\text{D.1}]$$

In the differentiated products case (see Question 1 for the homogeneous product case) firm 2's best response is $q_2 = A_2 - B_2 q_1$. Substituting this into firm 2's profit function takes care of the ‘max’ part of [D.1], so that [D.1] is equivalent to

$$\min_{q_1} \pi_2(q_1, A_2 - B_2 q_1) \quad [\text{D.2}]$$

Using the envelope theorem (Chapter 2, section I), the effect of an increase in firm 1's output on firm 2's profit, given that firm 2 makes its best profit maximizing response, is

$$d\pi_2/dq_1 = \pi_{21} + \pi_{22}(dq_2/dq_1) = \pi_{21} = -\gamma q_2 \quad [\text{D.3}]$$

Thus, since firm 2 cannot be forced to continue to produce if it would earn negative profits, firm 1 will wish to increase its output until firm 2's profit is zero. Solving $\pi_2(q_1, A_2 - B_2 q_1) = 0$ gives

$$q_1^x = (\alpha_2 - c_2)/\gamma \quad q_2^x = 0 \quad [\text{D.4}]$$

There are two potential difficulties for firm 1 with this minimax punishment. The first is that it may not be feasible. With $q_2 = 0$, the maximum amount that firm 1 can sell (by setting $p_1 = 0$) is

$$q_1^m = \alpha_1/\beta_1$$

which could be less than q_1^x . In the specific numerical example we are considering (see page 320) in fact $q_1^x = 18 > q_1^m = 10$. When the model's parameters are such that $q_1^x > q_1^m$ firm 1 will minimize firm 2's profit by producing its largest saleable output q_1^0 which satisfies

$$p^1 = \alpha_1 - \beta_1 q_1^0 - \gamma[A_2 - B_2 q_1^0] = 0 \quad [\text{D.5}]$$

But this raises the second difficulty: choosing q_1^0 and a price of zero leaves firm 1 with a loss. This difficulty can arise even if $q_1^x < q_1^m$. With $q_2 = 0$ and $q_1 = q_1^x$, firm 1 earns non-negative profits only if

$$p_1 - c_1 = \alpha_1 - \beta_1 q_1^x - c_1 \geq 0 \quad [\text{D.6}]$$

As the reader should check, by using [D.4] to substitute for q_1^x and referring back to Question 1 in section B, this condition need not be satisfied in the general linear differentiated products model. (It is obviously not in our numerical example.) If [D.6] does not hold and firm 1 wishes to break even whilst minimaxing firm 2 it must choose the output q_1^E satisfying

$$p_1 - c_1 = \alpha_1 - \beta_1 q_1^E - \gamma[A_2 - B_2 q_1^E] - c_1 = 0 \quad [\text{D.7}]$$

Note that in this case the firm being punished (firm 2) will be earning positive profits whilst the minimaxing firm 1 just breaks even. Punishment really does hurt more than being punished!

Thus there may be different feasible minimaxing outputs for the punishing firm depending on the parameters of the model and the loss we assume that the punishing firm can bear.

The definition of an *individually rational profit allocation* is straightforward. Let $\pi_i^c (= 0)$ be i 's profit when it is being minimaxed by q_j^x , and $\pi_i^E (> 0)$ that when it is being minimaxed by q_j^E . Then an individually rational profit allocation for firm i is any allocation which yields it a profit $\pi_i > \pi_i^c$ in the first case and $\pi_i > \pi_i^E$ in the second. Thus, use our numerical example and refer to Fig. 12.10(b). If when i is minimaxed it earns $\pi_i^c = 0$, then the set of individually rational profit pairs consists of all the points on and below the profit frontier and within the axes. If on the other hand i 's minimax profit is $\pi_i^E > 0$, as shown in the figure, then the individually rational profit pairs are all the points within the dotted lines drawn through π^E and on or below the profit frontier.

The Folk Theorem states that: *trigger strategies incorporating the punishment of being minimaxed forever can support all individually rational profit allocations as a Nash equilibrium for some set of values of the interest rate*. Thus, as compared to punishment by Cournot–Nash competition, minimax punishments considerably expand the set of possible

collusive outcomes (or equivalently, the set of interest rates at which a particular collusive outcome can be sustained).

The proof is similar to the Cournot–Nash punishment threat case. Let (π_1^*, π_2^*) be an individually rational profit pair which it is desired to sustain, and let $\pi_i^R \geq \pi_i^*$ again denote the one-period profit i can make by reneging and making its best response to j 's collusive output. The trigger strategies are: j will adhere to the agreement as long as i did in the previous period, but if i reneges in one period, j switches to the output $(q_j^R \text{ or } q_j^P)$ that maximizes i , in every subsequent period. It suffices to take the case in which i 's minimax profit is π_i^* .

Consider $t = 0$. If i reneges it gains $\pi_i^R - \pi_i^*$ now, but loses the infinite future stream $\pi_i^* - \pi_i^*$. This is strictly positive since π_i^* is individually rational. It will not renege if

$$\pi_i^R - \pi_i^* \leq (\pi_i^* - \pi_i^*)/r \quad [\text{D.8}]$$

If $\pi_i^* = \pi_i^P$ then the left-hand side is zero and the theorem certainly holds for all $r > 0$. If $\pi_i^* \neq \pi_i^P$ then we have the condition

$$r \leq (\pi_i^* - \pi_i^P)/(\pi_i^R - \pi_i^*) \quad [\text{D.9}]$$

Since the right-hand side is certainly positive, there always exists a range of interest rates for which this condition holds. Moreover, $\pi_i^* < \pi_i^P$ and so, comparing [D.9] and [C.12], the set of interest rates for which collusion can be sustained is clearly larger when minimax punishments are used.

The trigger strategies constitute a Nash equilibrium: if i believes j will play its trigger strategy then, given that condition [D.9] is satisfied, its best response is to play its trigger strategy and the result will be the collusive outcome (π_1^*, π_2^*) in every period. However, the minimax trigger strategies do not constitute a sub-game perfect equilibrium. To see this, suppose i has reneged at t , and consider the sub-game beginning at $t + 1$, $t = 0, 1, \dots$. If j minimizes i by producing q_j^R i makes its best response q_i^R , but q_i^R is not j 's best response to q_j^R (recall the only outputs that are mutually best responses are (q_1^*, q_2^*)). Therefore the output pair (q_i^R, q_j^R) is not a Nash equilibrium of the sub-game beginning at $t + 1$, in the contingency that i has reneged at t .

Thus we have arrived at the position: punishment by Cournot–Nash competition may not be very severe but is at least credible in the sense of sub-game perfection, while punishment by minimaxing may be sufficiently severe but is not credible.

The carrot-and-stick approach

D. Abreu has developed a simple but ingenious idea which allows more severe punishment than Cournot–Nash competition, but which also gives sub-game perfect strategies. Moreover, it dispenses with the assumption of eternal punishment, replacing it with the appealing idea that collusion would be resumed once a short sharp punishment for deviation has been inflicted. Collusion is sustained by the 'stick' of a profit-reducing output expansion to punish deviation and the 'carrot' of subsequent reversion to the collusive outputs, which plays the important role of inducing firms to accept the loss of profit required by the punishment phase.

We again denote the output and profit pairs that the firms choose as their collusive allocation by (q_1^*, q_2^*) and (π_1^*, π_2^*) respectively. These may or may not be on the profit

frontier. The strategies defined by Abreu are as follows: the firms produce the agreed output pair in each period as long as this was done in the previous period; if firm $i = 1, 2$ deviates in period t , the firms are to produce punishment outputs (q_1^P, q_2^P) in period $t + 1$, which in general depend on q_i^* and on which firm has deviated at t . Given the symmetry of our model, we can however consider only punishment outputs that do not depend upon which firm deviated at t .

If the firms produce punishment outputs at $t + 1$ then they revert to (q_1^*, q_2^*) at $t + 2$ and continue with these outputs unless one of them deviates ...; if firm $j = 1, 2$ deviates from its punishment output at $t + 1$, the punishment outputs (q_1^P, q_2^P) are again to be produced at $t + 2 \dots$; and so on. Deviation in the punishment phase (by either firm) results in *reimposition of punishment*, while acceptance of punishment (which is costly to both firms) results in reversion to collusion.

Denote punishment profits $\pi_i(q_1^P, q_2^P)$ by π_i^P . As before, we can consider the gains and losses to firm i from reneging on the agreed output q_i^* at $t = 0, 1, \dots$. Its immediate profit gain is $\pi_i^R - \pi_i^*$, where again $\pi_i^R = \pi_i(q_i^R, q_j^*)$ is the profit it makes from producing q_i^R , its best response to q_j^* . Assume that in the following period both firms do produce the punishment outputs so that i earns π_i^P (we will later justify this assumption of non-deviation during the punishment phase). If reneging at t is profitable, so will be reneging at $t + 2$, because the game is identical at every possible starting point. Thus firm i will renege at $t + 2$ and earn π_i^R , and then be punished at $t + 3$ and earn π_i^P and so on. The profit stream from reneging is the alternating infinite stream $\{\pi_i^R, \pi_i^P, \pi_i^R, \pi_i^P, \dots\}$, while that from not reneging is the constant infinite stream π_i^* , each of these streams beginning at t . Hence firm i will not renege at t if

$$\pi_i^R - \pi_i^* \leq \delta(\pi_i^* - \pi_i^P) + \delta^2(\pi_i^* - \pi_i^R) + \delta^3(\pi_i^* - \pi_i^P) + \dots \quad i = 1, 2 \quad [\text{D.10}]$$

where the left-hand side is the immediate gain from reneging and the right-hand side is the discounted value at t of the difference in profit streams from not reneging and reneging. Notice there is no guarantee that this right-hand side is even positive, let alone that it exceeds the left-hand side, since the $\pi_i^* - \pi_i^R$ terms are all negative. Now,

$$\sum_{t \in E} \delta^t = (1 - \delta^2)^{-1} \quad E = \{0, 2, 4, \dots\}$$

and

$$\sum_{t \in D} \delta^t = \delta(1 - \delta^2)^{-1} \quad D = \{1, 3, 5, \dots\}$$

(see Question 5, Exercise 12C), so we can rearrange [D.10] to obtain the condition,

$$\pi_i^R - \pi_i^* \leq \delta(\pi_i^* - \pi_i^P) \quad i = 1, 2 \quad [\text{D.11}]$$

or

$$r \leq \frac{\pi_i^* - \pi_i^P}{\pi_i^R - \pi_i^*} - 1 \quad i = 1, 2 \quad [\text{D.12}]$$

Thus, since $r > 0$, a necessary (but not sufficient) condition for a collusive profit π_i^* to be sustainable (given that there is no deviation in the punishment phase) is that

$$\pi_i^R - \pi_i^* \leq \pi_i^* - \pi_i^P \quad i = 1, 2 \quad [\text{D.13}]$$

That is, that there exist sufficiently large outputs q_i^P to generate sufficiently small profit π_i^P that the one-period gain from reneging can be offset by a one-period punishment. Whether this will hold depends on the market structure – the demand and cost functions – which determine the relationships among these profit values. If condition [D.12] is satisfied, then it is clearly in the firms' interests to choose π_i^P so as to equate the right-hand side with r , since the larger is π_i^P , the smaller the loss of profit in the punishment phase. On the other hand for the smallest possible π_i^P feasible in the market, [D.12] defines the highest interest rate for which collusion is sustainable.

All this assumes that there is no defection in the punishment phase. Consider the sub-game beginning at $t = 1, 2, \dots$ when a firm has reneged at $t - 1$. Firm $i = 1, 2$ is to produce q_i^P and earn π_i^P . If it does this, and condition [D.12] is satisfied, so that collusion from $t + 1$ onward will be maintained, then it earns the profit stream consisting of π_i^P at t and π_i^* from $t + 1$ onward. This has a discounted value at t of $\pi_i^P + (\pi_i^*/r)$. Let q_i^{RP} denote i 's best response output to q_j^P and $\pi_i^{RP} = \pi_i(q_i^{RP}, q_j^P)$ the profit it will make if it reneges in the punishment phase in period t . If it reneges at t , under the strategy described above punishment is reimposed at $t + 1$. But if it pays to renege at t it will pay to renege also at $t + 1$, and in every future period when punishment is reimposed. Thus associated with the decision to renege in the punishment phase is the infinite stream of profit consisting of π_i^{RP} forever. This has a discounted value at t of $\pi_i^{RP} + (\pi_i^{RP}/r)$. Thus, for i to keep to the agreement in the punishment phase and *not* renege we require

$$\pi_i^P + (\pi_i^*/r) \geq \pi_i^{RP} + (\pi_i^{RP}/r) \quad i = 1, 2 \quad [\text{D.14}]$$

or

$$\pi_i^{RP} - \pi_i^P \leq (\pi_i^* - \pi_i^{RP})/r \quad i = 1, 2 \quad [\text{D.15}]$$

Thus the one period gain from reneging in the punishment phase must be more than offset by the present value of loss of profit resulting from having the punishment phase continually reimposed rather than restoring collusion. We can express this in terms of the interest rate

$$r \leq \frac{\pi_i^* - \pi_i^{RP}}{\pi_i^{RP} - \pi_i^P} \quad i = 1, 2 \quad [\text{D.16}]$$

If conditions [D.12] and [D.16] are then *simultaneously* satisfied, the profit pair (π_1^*, π_2^*) is sustainable by the strategies described. Note that the two conditions are mutually reinforcing and must hold simultaneously. [D.12] ensures that it never pays to deviate given that punishment will be inflicted; [D.16] guarantees that punishment will be inflicted (even though it hurts the punisher) given that in the period following punishment collusion will be reinstated and maintained.

We can show that the strategies are sub-game perfect equilibrium strategies, so that the threats inherent in them are credible on this criterion. There are four kinds of sub-game:

1. The game itself, beginning at $t = 0$. In this, if i expects j to adhere to the specified strategy then, given that conditions [D.12] and [D.16] are satisfied, its best response is also to adhere, and so the strategies are a Nash equilibrium for the entire game.
2. A (proper) sub-game beginning at $t = 1, 2, \dots$, in which nobody has reneged at $t - 1$. Since this game is identical to the game at $t = 0$, the strategies are a Nash equilibrium for these sub-games.

3. A (proper) sub-game beginning at $t = 1, 2, \dots$, in which one firm reneged at $t - 1$. Given conditions [D.15] and [D.16], i 's best response to q_j^P at $t + 1, \dots$, is itself to produce q_i^P at t and q_i^* at $t + 1, \dots$. Thus the strategies induce a Nash equilibrium in these sub-games.
4. A (proper) sub-game beginning at $t = 2, \dots$, in which one firm reneged at $t - 2$ and (q_1^P, p_2^P) was produced at $t - 1$. Since this game is identical to the game beginning at $t = 0$, we again have a Nash equilibrium.

Thus, since the strategies induce Nash equilibria in all possible sub-games they are sub-game perfect equilibrium strategies.

We can illustrate the strategies using the numerical example on which Fig. 12.10(b) is based and consider whether the joint-profit maximizing profit pair (π_1^*, π_2^*) can be sustained. The symmetry in this example greatly simplifies calculations.

Table 12.3 sets out the results. We take three possible interest rates: 2 per cent, 5 per cent and 25 per cent respectively. These could be thought of as showing the implications of different period lengths: since a reasonable *annual* interest rate is 25 per cent, the first case corresponds roughly to a period of a month, the second to a period of one quarter and the third to one year. The length of the period determines the duration of punishment as well as the length of time for which a firm can make profits from reneging before retaliation takes place.

Table 12.3

	$r = 0.02$	$r = 0.05$	$r = 0.25$
$\pi_i^P = \pi_i^* - (1+r)[\pi_i^* - \pi_i^P]$	12.9263	12.9094	12.7969
$q^P = (q + [81 - 6\pi_i^*]^{1/2})/3$	3.6184	3.6274	3.6847
$q_i^{RP} = 4.5 - 0.25q^P$	3.5954	3.5032	3.5788
$\pi_i^{RP} = (q - 0.5q^P)q_i^{RP} - (q_i^{RP})^2$	12.9270	12.9107	12.8077
$(\pi_i^* - \pi_i^{RP})/(\pi_i^{RP} - \pi_i^P)$	13.5 - 12.9270 12.9270 - 12.9263	13.5 - 12.9107 12.9107 - 12.9094	13.5 - 12.8077 12.8077 - 12.7969

Given the assumed interest rate, the first line of the table uses [D.12] as an equality to calculate the *minimum* required punishment profit π_i^P – recall from Table 12.2 that $q_i^m = 3$, $\pi_i^m = 13.5$, and so π_i^R is given by $q_i^R = A_i - B_i q_i^m$. We next calculate the punishment outputs required to generate π_i^P . For simplicity we assume *symmetric punishments*: the firms are assumed to produce the same outputs q^P in the punishment phase. Solving $\pi_i^P = (\alpha - c - \gamma q^P)q^P - (q^P)^2$ for q^P yields two roots but we take that root which exceeds $q_i^m = 3$. The output that i will choose in the punishment phase if it reneges, q_i^{RP} , is found from the best response function, $q_i^{RP} = A_i - B_i q^P$. This gives i 's profit from reneging in the punishment phase, $\pi_i^{RP} = \pi_i(q_i^{RP}, q^P)$. The last line of the table gives the value of the right-hand side of [D.16]. If the implied value is greater than r , then the joint profit maximizing allocation is sustainable by the punishment output q^P . In each case in the table this is true by a very substantial margin. In fact, for any value of r up to about 2.8, i.e. an interest rate of 280 per cent, this allocation is sustainable in this example.

The reason for the ease with which collusion can be sustained in this example should

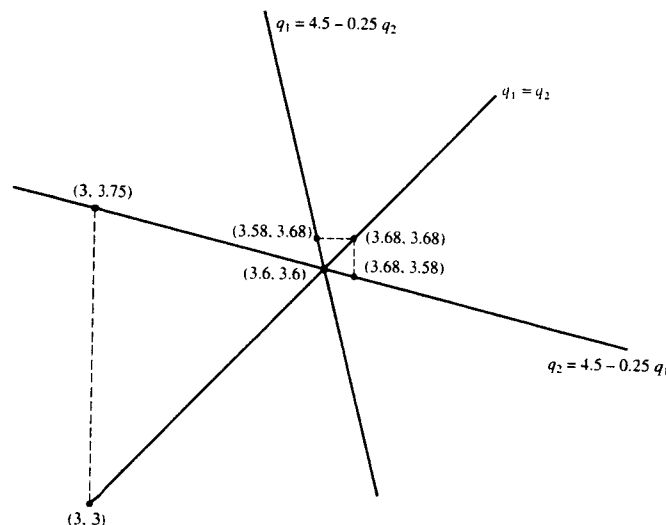


Fig. 12.13

be clear from the last line of the table: the gain in profit from reneging on the punishment output is very small relative to the difference between the profit from colluding and that earned by reneging on punishment. Fig. 12.13 illustrates the various output pairs in this example. The (symmetric) collusive output pair $(3, 3)$ is of course below the Cournot–Nash equilibrium pair $(3.6, 3.6)$. Suppose firm 2 reneges by producing its best response output 3.75. From Table 12.3 we see that the most profitable symmetric punishment pair (which gives π_i^P satisfying [D.12] as an equality) is (for $r = 0.25$), $(3.68, 3.68)$. This represents more severe punishment than Cournot–Nash competition. If either firm reneges on the punishment output, it will produce $q_i^{RP} = 3.58$. But we know from condition [D.31] that it is more profitable for each firm to accept the lower profit π_i^P in the punishment period and then return to the collusive profit than to have punishment reimposed.

Note that in the punishment phase not only does the punisher have an incentive to bear the costs of inflicting the punishment, but the firm being punished has an incentive to accept or ‘cooperate in’ its punishment – both firms produce their punishment outputs q^P . Moreover, if firm 2 reneges at t , and then 1 does not inflict the punishment at $t + 1$, the strategies call for 2 to punish 1 at $t + 2$: thus the cheat punishes the non-cheat for not punishing him for cheating! Nevertheless, as long as our criterion of credibility is sub-game perfection, the threats underlying the strategies are credible and the equilibrium output path, when [D.12] and [D.16] are satisfied, will be the agreed outputs (q_1^*, q_2^*) ($= (q_1^m, q_2^m)$ in our example) in every period.

Whether any given profit pair can be sustained by Abreu’s punishment strategies depends on the cost and demand functions and the interest rate, since the former determine the profit functions and the latter the present value of future losses from reneging. A further analysis of the stick and carrot approach by Fudenberg and Maskin (1986) shows that

any individually rational profit pair can be sustained for some range of interest rates as a sub-game perfect equilibrium. They base the punishment outputs on *mutual minimaxing* by the two firms: each firm produces the output it would use to minimax the other. The resulting profit for the deviating firm is therefore in general worse than it would be in the case envisaged in the original Folk Theorem. Consequently there always exists a duration of the period of punishment such that any individually rational profit allocation can be sustained by credible threats of this type of punishment, for some set of interest rates. However, punishment by mutual minimaxing is only guaranteed to work in the case of two firms. Whereas all the propositions on sustainability of collusion we have so far considered generalize to more than two firms, the Fudenberg–Maskin result does not. For example, in the case of three firms, there may not in general exist output levels q_1^* , q_2^* , q_3^* , such that q_i^* and q_j^* simultaneously minimax firm k ($i, j, k = 1, 2, 3$, $i \neq j$, $j \neq k$, $i \neq k$).

Renegotiation proof strategies

The criterion of the credibility of threats of punishment we have applied so far is sub-game perfection. The underlying idea of the trigger strategies considered so far is that the firms meet, agree on a collusive equilibrium (q_1^*, q_2^*) , agree also on trigger strategies (say of Abreu’s type), and then spend the rest of time independently implementing those strategies. It is as if they conclude a contract, which is self enforcing because of the sub-game perfect punishment threats it contains, and which is never re-negotiated. Since the initial ‘contract’ was self-enforcing rather than legally binding the parties cannot include in it a tacit agreement not to renegotiate by introducing a suitable punishment for proposing renegotiation: any renegotiation proposal can also cover the punishment for making the proposal. It is impossible for the parties to commit themselves not to renegotiate: if such commitment is possible they would presumably have committed themselves to the collusive solution at the start, without the need for elaborate punishment strategies. Suppose then that firm i has reneged at time t . Firm j now faces the prospect of punishing i in the knowledge that if it does not, under the agreed (Abreu) strategy firm i will punish firm j next period. What is then to stop i suggesting to j that instead of going through this self-lacerating process they renegotiate the agreement and simply start colluding again? Certainly, at this point renegotiating to a collusive outcome is better for both i and j . But if this is anticipated *ex ante* the Abreu strategies no longer embody credible threats.

The requirement that strategies be credible in the sense that they are renegotiation proof in general reduces the set of possible equilibrium allocations to a subset of those that can be supported as sub-game perfect equilibria – it represents a further refinement of Nash equilibrium. To see what that is, suppose the firms wish to support a profit pair (π_1^*, π_2^*) . If firm i is called upon to punish firm j after a deviation, then it must be in its interests to do so rather than allow itself to be ‘negotiated back’ to the profit pair (π_1^*, π_2^*) . This will be assured if i ’s profit in the punishment phase, π_i^P , satisfies

$$\pi_i^P \geq \pi_i^* \quad [\text{D.17}]$$

But in addition, for the punishment to be credible, the punishment pair (π_1^P, π_2^P) must itself be renegotiation proof. Thus an allocation (π_1^*, π_2^*) is *weakly renegotiation proof* if there is a weakly renegotiation proof profit pair (π_1^P, π_2^P) which

- (a) satisfies conditions [D.12] and [D.16] so that the punishment strategy is sub-game perfect;

an allocation which is better for the punisher and worse for the cheat than the initially agreed-upon allocation. In its strong form, this condition implies that the set of equilibrium profit allocations is a subset of the points on the market profit frontier.

The theories surveyed in this chapter provide testable hypotheses concerning the conditions under which we would observe collusion: the conditions involve measurable parameters of the firms' cost and demand functions and the interest rates the firms face. A drawback is that there still remains a wide set of possible equilibrium allocations: theories concerning how equilibria may be supported by threats do not predict a unique market equilibrium. Further hypotheses are required to predict which allocation in the set of sustainable allocations will actually be the market outcome. This is still an open question in oligopoly theory.

Most of the field of economics known as industrial organization is taken up by the theoretical and empirical study of oligopolistic markets. Here we have considered just some of the central themes in this literature: hypotheses about equilibrium resource allocation for homogeneous and differentiated products when the market is viewed as a one-shot game; and the conditions under which firms may make and sustain cooperative agreements when there is repeated market interaction between the firms. In general the restriction to only two firms was purely simplifying in the sense that it is relatively straightforward to extend the solutions of the various models to the case of $n > 2$ firms. On the other hand, many interesting aspects of oligopolistic markets have not been considered, and the reader is urged to follow up at least some of the references at the end of this chapter, and further to study this most important type of market.

Notes

1. Moreover, each player knows that the other knows this, knows that the other knows he knows, and so on to any required degree. This is referred to as the *common knowledge* assumption. If it is not made we have a different game and there may be other interesting strategic possibilities. Consider, for example, a poker game in which player *A* has a weak hand, *B* knows this (because the cards are marked) but *A* does not know he knows. Then *A* may try to bluff by betting heavily as if he had a strong hand, in a way he would not if he knew *B* knew his hand. What do you think might happen if *A* knows the cards are marked, but *B* does not know he knows that?
2. With just two firms of course perfect competition would not be feasible especially if outputs are not homogeneous. This 'price = marginal cost' allocation is simply a useful benchmark against which to assess the various duopoly outcomes, since it corresponds to the allocatively efficient outcome.
3. It can be shown that collusion can also be sustained in finitely repeated games if there are multiple Nash equilibria of the constituent game, or if each firm believes there is some probability, however small, that the other firm is of the type that would choose the collectively rational strategy – collude – rather than the individually rational strategy, cheat. See the references at the end of this chapter.
4. In actually deriving the profit frontier it is more useful to approach the solution to the problem in [C.4] as follows. For $\lambda \in (0, 1)$ solve the *unconstrained* problem, with parameter λ

$$\max_{q_1, q_2} \lambda \pi^1(q_1, q_2) + (1 - \lambda) \pi^2(q_1, q_2)$$

The solutions (q_1^*, q_2^*) are functions of λ . Varying λ over the interval $[0.25, 0.75]$ then generates numerically the solution pairs in Fig. 12.10(a).

References and further reading

Serious study of oligopoly requires extensive knowledge of game theory. A selective bibliography is:

- R. D. Luce and D. Raiffa. *Games and Decisions*, John Wiley, New York, 1966.
 H. Moulin. *Game Theory for the Social Sciences* (2nd edn), New York University Press, New York, 1986.
 J. Friedman. *Game Theory with Applications to Economics*, Oxford University Press, Oxford, 1986.
 D. Fudenberg and J. Tirole. *Game Theory*, MIT Press, 1991.

On the economics, the discussion here of mixed strategy equilibrium in the Edgeworth model is based on:

- R. Levitan and M. Shubik. 'Price duopoly and capacity constraints', *International Economic Review*, 13, 1972, 111–22.
 D. Kreps and J. Scheinkman. 'Quantity precommitment and Bertrand competition yield Cournot outcomes', *Bell Journal of Economics*, 14, 1983, 326–37.
 C. Davidson and R. Deneckere. 'Long-term competition in capacity, short-run competition in price, and the Cournot model', *Rand Journal of Economics*, 17, 1986, 404–15.

The discussion of trigger strategies was based on:

- D. Abreu. 'Extremal equilibria of oligopolistic supergames', *Journal of Economic Theory*, 39, 1986, 191–235.
 J. Farrell and E. Maskin. 'Renegotiation in Repeated Games', *Games and Economic Behaviour*, 1, 1989, 327–60.
 D. Fudenberg and E. Maskin. 'The Folk Theorem in repeated games with discounting or with incomplete information', *Econometrica*, 54, 1986, 533–56.

On oligopoly in general, very comprehensive treatments are given by:

- J. Tirole. *The Theory of Industrial Organization*, MIT Press, 1988.
 S. Martin. *Advanced Industrial Economics*, Blackwell, Oxford, 1992.
 J. Friedman. *Oligopoly and the Theory of Games*, North-Holland, Amsterdam, 1977.

and all the chapters in Vol 1, Part 2, of

- R. Schmalensee and R. Willig (eds) *Handbook of Industrial Organization*, Elsevier Science Publishers, 1989.

CHAPTER 13

Alternative theories of the firm

A. Introduction

In the theory of the firm set out in earlier chapters the firm was assumed to choose inputs, outputs and other decision variables to maximize its profits. Since the profits of the firm accrue to its owners, this specification of the firm's objectives rests on two tacit assumptions. The first is that the only aspect of the firm which the owners care about is the profit its activities generate. The second is that the owners can control the firm's activities and ensure that they are indeed profit maximizing. Both of these assumptions can be questioned and in this chapter we consider three theories of firm behaviour resulting from different assumptions about the owners' preferences or their ability to control the firm.

In section B the single owner of the firm cares about the income he gets from the firm and the amount of effort he supplies to it. The section shows that in the absence of a market in effort, utility maximization by the owner implies profit maximization by the firm only if the owner's preferences are of a special type. In section C the owner is only concerned with her income from the firm which is run for her by a manager. We investigate the circumstances in which the delegation of control from the owner (principal) to the manager (agent) results in the firm not maximizing profit. We consider how the owner can design contracts with her manager to mitigate the effect of the manager's discretion on her income from the firm. The simple delegated choice model of the firm set out in section C is a useful introduction to the more complete treatment of principal-agent models in Chapter 22 section E. In section D the firm is a workers' cooperative or partnership. It is owned by its workers who supply labour and in exchange receive a share of its revenue after deduction of all non-labour costs.

B. The entrepreneurial firm

It has sometimes been argued that the axiom of profit maximization has more plausibility when it refers to owner-controlled, or *entrepreneurial*, firms. The owner receives the firm's profit as income, and hence to maximize his profit is to maximize his income, a

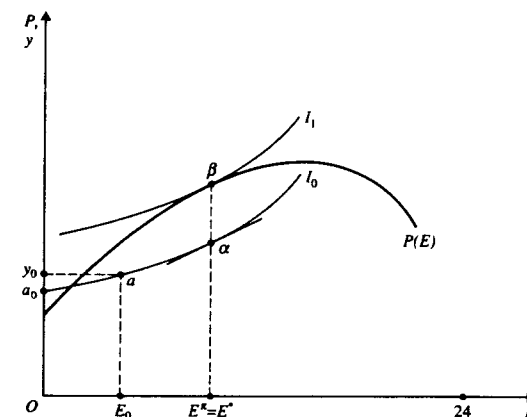


Fig. 13.1

reasonable-sounding aim. T. Scitovsky, however, argued that matters are not so straightforward. In generating the firm's 'gross income' (defined as revenue less all costs *except that of his own services*), the entrepreneur must expend time and effort. If we regard the amount of effort as constant per unit time, then we can measure the entrepreneur's input in terms of, say, the number of hours per day, up to a maximum of 24, he devotes to the work of the firm. We expect that, at least over some range, the firm's gross income increases with the entrepreneurial input as he devotes more time to getting business and controlling costs, but at a diminishing rate. The curve $P(E)$ in Fig. 13.1 shows the relation between gross income per day measured on the vertical axis, and his input, measured in hours per day, along the horizontal (cf. the analysis in Chapter 5, section C).

The fact that $P(E)$ has a positive intercept indicates that even if the entrepreneur put no effort into his firm it would still generate an income for him. If the entrepreneur had to devote at least some effort to the firm in order to get a positive gross income, the intercept would be negative. The value of $P(0)$ does not affect the analysis, provided that over some range of E the firm produces a positive income for the entrepreneur.

In defining gross income we did not make any allowance for the cost of the entrepreneur's own services, so that the curve $P(E)$ does not measure profit, which was defined in earlier chapters as revenue minus *all* opportunity costs. We therefore need a measure of the opportunity cost of the services the entrepreneur provides to his firm. By providing his services to the firm the entrepreneur foregoes the opportunities of using them elsewhere to generate income or of having all his time available for leisure. To measure the value of these foregone alternatives we must make some assumptions about his preferences as regards income and effort.

It is plausible that these preferences are similar to those of the labour suppliers examined in Chapter 5, section C. Letting $u(E, y)$ be the entrepreneur's strictly quasi-concave utility function, we would expect that he prefers less effort to more ($\partial u / \partial E = u_E < 0$) and more income to less ($\partial u / \partial y = u_y > 0$). Fig. 13.1 shows one possible indifference map for the

entrepreneur. The indifference curves are upward sloping because he must be compensated with additional income if he supplies additional effort. The slope of an indifference curve increases as he supplies more effort, indicating that additional effort becomes increasingly distasteful and must be compensated for by larger increases in income.

Suppose that the entrepreneur's best alternative to supplying effort to run his firm is the effort-income combination E_0, y_0 at a in Fig. 13.1. The indifference curve I_0 shows all effort-income combinations which yield the same utility as a .

$$u(E, y) = u(E_0, y_0) = u^0 \quad [\text{B.1}]$$

I_0 shows the minimum income that he must be given to induce him to supply different levels of effort if he is to be no worse off than in his foregone alternative. The height of I_0 is a monetary measure of the opportunity cost to him of supplying effort to his firm and thereby foregoing the alternative a . Solving [B.1] for y , as a function of E and the utility level achieved in the alternative a , gives the opportunity cost A of effort as $A(E; u^0)$. Writing [B.1] as an implicit function and using the implicit function rule, confirms that the marginal opportunity cost of effort is the slope of the indifference curve I_0 :

$$\frac{\partial A(E; u^0)}{\partial E} = A_E(E; u^0) = \frac{dy}{dE} \Big|_{u=u^0} = -\frac{u_E}{u_Y} \quad [\text{B.2}]$$

Two points should be noted about the opportunity cost of effort. First, it depends on the preferences of the entrepreneur and is therefore unobservable and likely to differ for different individuals. Thus profit maximization may imply different effort levels for different entrepreneurial firms, even if the entrepreneurs confront the same $P(E)$ curve. Second, in general, the marginal opportunity cost of effort depends on the level of utility achieved in the foregone alternative. As we will see, this is crucial in answering the question of whether the entrepreneurial firm will in fact maximize profit.

Since the height of the indifference curve I_0 measures the opportunity cost of the effort the entrepreneur supplies to his firm, the firm's profit in Fig. 13.1 at each effort level is the vertical distance between the income curve $P(E)$ and I_0 :

$$\Pi(E) = P(E) - A(E; u^0) \quad [\text{B.3}]$$

Profit is maximized at the effort level E^* , where the slope of $P(E)$ at β is equal to the slope of I_0 at α :

$$P'(E^*) = A_E(E^*; u^0) \quad [\text{B.4}]$$

But the entrepreneur chooses his effort level to maximize his utility $u(E, y)$, subject to the constraint $y = P(E)$. (As Question 1, Exercise 13B asks you to demonstrate, the analysis is not substantively different if he has an endowed income \bar{y} , so that his income constraint is $y = P(E) + \bar{y}$.) Does his utility maximizing effort E^* also maximize the firm's profit: is $E^* = E^*$? Substituting the constraint into the utility function, the entrepreneur's problem is

$$\max_E u(E, P(E)) \quad [\text{B.5}]$$

and, assuming a non-corner solution, his optimal effort level E^* satisfies the first-order

condition

$$u_E(E^*, P(E^*)) + u_Y(E^*, P(E^*))P' = 0 \quad [\text{B.6}]$$

which can be rearranged to get

$$P'(E^*) = -\frac{u_E(E^*, P(E^*))}{u_Y(E^*, P(E^*))} \quad [\text{B.7}]$$

The left-hand-side of [B.7] is the slope of the $P(E)$ curve and the right-hand side is the slope of an indifference curve at the optimal point. In terms of Fig. 13.1, he maximizes his utility by moving along $P(E)$ until he reaches the highest possible indifference curve. The optimum effort level E^* is where an indifference curve is tangent to $P(E)$. In Fig. 13.1 this is at β , where I_1 is tangent to $P(E)$ at the effort level $E^* = E^*$. In this case utility maximization by the entrepreneur leads to profit maximization.

However, in general, the tangency of $P(E)$ with an indifference curve need not occur at E^* . From [B.4] and [B.2], the profit maximizing effort level E^* is defined by the slope of P being equal to the slope of the foregone opportunity indifference curve I_0 , whereas the utility maximizing effort level E^* is defined by the slope of P being equal to the slope of an indifference curve which can be reached by moving along P . Thus E^* and E^* will coincide only if indifference curves are vertically parallel: their slope must depend only on the level of effort and not on the income level. In Fig. 13.1 the preferences of the entrepreneur satisfy this rather special requirement: the slope of I_1 is equal to the slope of I_0 at all levels of E . Since the slope of the indifference curves measures the marginal cost of effort, vertically parallel indifference curves mean that the marginal cost of effort depends only on effort. Thus a businessman who, as he gets richer, spends more time on the golf course, even though the marginal return to his effort is unchanged, is not a profit maximizer.

Entrepreneurial input market

Suppose that there is a competitive market on which it is possible to buy and sell entrepreneurial inputs at a given price w . The entrepreneur can choose to run his own business, employ someone else to run it, or become employed running a firm for someone else. Suppose also that he does not mind whether he is 'his own boss' or whether he works for someone else for the same income. In these circumstances the Scitovsky conclusion that profit maximization requires a special type of preference is no longer valid. The reason is that, with an entrepreneurial input market, the opportunity cost of the entrepreneur's effort is not the sum of money required to induce him to work in the firm but rather the sum he could get if he sold his services on the market, rather than supplying them to his own firm.

The existence of the entrepreneurial input market enables the entrepreneur to separate the decisions on how much effort should be used in his firm and how much effort he should supply. Let E_f denote the level of entrepreneurial input used in his firm and E be the amount of effort that he supplies on the entrepreneurial input market.

The existence of the entrepreneurial input market means that the opportunity cost of

effort used in the firm is wE_f : the sum the entrepreneur would have to pay someone to work for him. Thus the firm's profit is

$$\Pi(E_f) = P(E_f) - wE_f \quad [\text{B.8}]$$

and his income is

$$y = \Pi(E_f) + wE \quad [\text{B.9}]$$

The entrepreneur chooses E and E_f to solve

$$\max_{E, E_f} u(E, \Pi(E_f) + wE) \quad [\text{B.10}]$$

The first-order conditions, assuming a non-corner solution, are

$$\frac{du}{dE} = u_E + u_y w = 0 \quad [\text{B.11}]$$

$$\frac{du}{dE_f} = u_y [P'(E_f) - w] = 0 \quad [\text{B.12}]$$

Rearranging [B.11] gives the condition for his optimal own effort supply:

$$-u_E/u_y = w \quad [\text{B.13}]$$

In supplying effort he acts just like the labour suppliers in Chapter 5, section C, equating his marginal cost of effort to the marginal increase in income from selling extra effort. Since $u_y > 0$, [B.12] implies that the optimal entrepreneurial input into the firm satisfies

$$P'(E_f^*) = w \quad [\text{B.14}]$$

so that the level of effort used in the firm is profit maximizing. The entrepreneur's income is $y = \Pi(E_f) + wE$ and, because his utility $u(E, \Pi(E_f) + wE)$ depends on the level of effort employed in the firm only via its effect on the firm's profit, he will want to make the profit from the firm as large as possible.

The solution is illustrated in Fig. 13.2. $P(E_f)$ shows the gross income from the entrepreneur's firm as a function of the managerial effort employed in it. The line OW has slope equal to the price of entrepreneurial effort w and measures the opportunity cost of effort wE_f used in the firm. The firm's profit is the vertical distance between OW and $P(E_f)$ and is maximized at E_f^* . If the entrepreneur decided to supply this amount of effort to his firm ($E = E_f^*$) he would have the income-effort combination β on $P(E_f)$. Consider the line W^*W^* which has slope w and is tangent to $P(E_f)$ at β . If he decides to put $E = E_f^*$ of his effort into his firm and also to sell some of his effort on the entrepreneurial labour market he would move rightward up the line W^*W^* . On the other hand, he could decide to reduce his effort, whilst keeping the amount of effort in the firm constant at E_f^* by buying effort from the market equal to $E_f^* - E$. He would then move leftward down W^*W^* . Thus the line W^*W^* is the constraint along which he can transact in the managerial labour market, given that he has fixed the input into his firm at the profit maximizing level E_f^* . He will choose a point on W^*W^* which gives him the largest utility. Notice that if he had fixed E_f at some other level which did not maximize profit, the labour market opportunity line along which he could transact would lie below W^*W^* . For example with

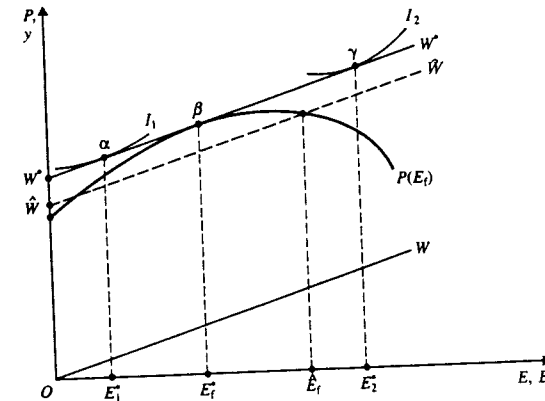


Fig. 13.2

$E_f = \hat{E}_f$, he would be on the market opportunity line $\hat{W}\hat{W}$. Since he can achieve a higher indifference curve on W^*W^* than on $\hat{W}\hat{W}$, he will wish to maximize the profit from his firm.

The entrepreneur's optimal position on W^*W^* depends on his preferences. There are three possible types of solution:

1. An indifference curve like I_1 is tangent to W^*W^* at α . He supplies effort E_1^* and sets the level of effort used in his firm at $E_f = E_f^*$. This solution could be achieved by putting E_1^* of his effort into the firm and then buying in $E_f^* - E_1^*$ from the labour market. Equivalently, he could buy in all the effort required for his firm and sell E_1^* of his labour on the market. We have assumed that he is indifferent to whether he works in his own firm or for someone else, so the model serves only to predict the level of input into his firm and his total effort supply, not how he divides his effort between working for himself and for others.
2. An indifference curve like I_2 is tangent to W^*W^* at γ . He sets $E_f = E_f^*$ and supplies E_2^* of his effort. This solution could be achieved by supplying E_f^* of his own effort to his firm and then selling $E_2^* - E_f^*$ of his effort on the market.
3. An indifference curve (not shown) is tangent to W^*W^* at β . This solution could be achieved by his supplying all the effort required for his firm and neither buying nor selling labour in the market.

The crucial result of this analysis is the *separation* between the production decision of the entrepreneur in his role as owner of the firm and the decision of the entrepreneur in his role as effort supplier. (This separation appears again in Chapter 15 when we examine investment and consumption choices, which also involve a decision-maker who has *both production and exchange opportunities*.) The production decision (E_f) depends on the productive opportunities embodied in $P(E)$ and on the market price of effort w . The

preferences of the entrepreneur as regards income and his effort have no influence on production decision and affect only his effort supply.

The existence of the market in entrepreneurial effort establishes an objective opportunity cost of effort used in the firm and enables the entrepreneur to separate his production decision from his effort supply decision. The Scitovsky conclusion no longer holds: profit maximization is not a special case.

Exercise 13B

- How is the entrepreneur's decision altered if he has an endowed income \bar{y} which he gets irrespective of whether he supplies effort to his firm or elsewhere?
- Show that if the entrepreneur has the *quasi-linear* utility function $u = g(E) + y$, with $g' < 0$, $g'' < 0$, his effort choice will always maximize profit.
- Suppose that the entrepreneur's concave production function is $q = F(E, z)$, where z is an input which he buys at a constant price p_z and q is output which he sells at a constant price p . Assume there is no market for his effort.
 - Derive the entrepreneurial firm's output supply and input demand functions.
 - Are the entrepreneur's input and output responses to changes in p and p_z different from those of the competitive firm examined in earlier chapters?
 - Describe the long-run equilibrium if there is free entry into the entrepreneur's industry.
- What are the consequences of assuming that, although the entrepreneur dislikes supplying effort, other things being equal he prefers to work for himself rather than for someone else?
- * A physician has a number of 'private' patients, whom he can arrange in order of fee per minute spent in attendance, from highest to lowest. He may also work in the state health service, at a given fee per unit time. Adapting Fig. 13.2 to this case, state necessary and sufficient conditions under which he would attend private patients and work for the state health service. Suppose that he is now forced to choose *either* to work privately *or* for the state health service. Analyse the determinants of his choice, indicating also the effects on his income and total supply of effort. Finally, suppose that a special tax is levied on his earnings from private practice, and analyse the consequences.

C. Agency theory and the separation of ownership from control*

The introduction of a market in managerial services raises the issue of the separation of ownership from control, discussed in Chapter 6. Suppose that the owner of the firm, denoted by O , must delegate management of the firm entirely to a professional manager, A . This may be because O does not possess the type of skills necessary to run the firm or because she prefers to use her labour in other activities. Since O does not supply any effort to her firm, she will be only concerned with the income it yields her. For the moment we retain the simple structure of the model of the previous section, and reinterpret Fig. 13.1

ignore I_1 from now on) as follows. Let E now be the manager's effort supply, the function $P(E)$ be the profits of the firm *before* subtraction of the manager's pay, and I_0 be the manager's *reservation indifference curve*. If we measure the manager's pay, y , as well as profit, on the vertical axis, I_0 shows the set of (E, y) pairs such that the manager is indifferent between working for the firm and taking his next best employment opportunity. We assume that the manager's utility function is quasi-linear: $u(E, y) = v(y) - E$, where $v' > 0$, $v'' < 0$. (u is no longer an ordinal utility function, but a von Neumann–Morgenstern cardinal utility function. See Chapter 19 for details.)

A 's reservation utility level u^0 is the utility A derives from his next best employment. All effort–income combinations along the reservation indifference curve I_0 yield u^0 to A . We assume that O can hold A down to his reservation utility level. (Question 1, Exercise 13C, considers an alternative assumption.) Thus the height of I_0 measures the cost of A 's effort to O : the minimum amount A must be paid to induce him to supply effort for O . Setting

$$v(y) - E = u^0 \quad [\text{C.1}]$$

and solving for y in terms of E , we can write

$$y = v^{-1}(u^0 + E) = y(E) \quad [\text{C.2}]$$

The graph of this function is the curve I_0 in Fig. 13.1. Differentiating through [C.2] we have

$$y'(E) = dy/dE = 1/v'(y) \quad [\text{C.3}]$$

as the slope of I_0 , or the *marginal cost* to the owner of the manager's effort, at any E .

O 's income from the firm is the profit of the firm: the difference between the gross profit and the amount she must pay A to supply effort. O will want to set $E = E^*$ in Fig. 13.1, since this is the solution to the profit maximization problem

$$\max_E P(E) - y(E) \Rightarrow P'(E^*) = y'(E^*)$$

If the owner:

- knows the $P(E)$ and $y(E)$ functions; and
- observes the gross profit outcome $P(E)$, or E itself,

then no problem arises out of the separation of ownership from control. She can engage A on a contract which specifies the payment $y^* = y(E^*)$ conditional on the profit outcome being $P(E^*)$ (or equivalently on A putting in effort E^*), and which pays $\bar{y} < a_0$ (refer to Fig. 13.1) if any outcome other than $P(E^*)$ results. Under this contract A does best by supplying E^* , since any lower E can be costlessly detected and results in lower utility than $v(y^*) - E^* = u^0$.

When can the separation of ownership from control present difficulties to O ? It seems natural to suppose that O will always be able to observe the gross profit outcome $P(E)$, and so, as long as there is a one-to-one relationship between P and E , the owner can always *infer* the value of E and apply the penalty clause for a deviation. We can begin to construct a model in which the separation of ownership from control could cause a problem

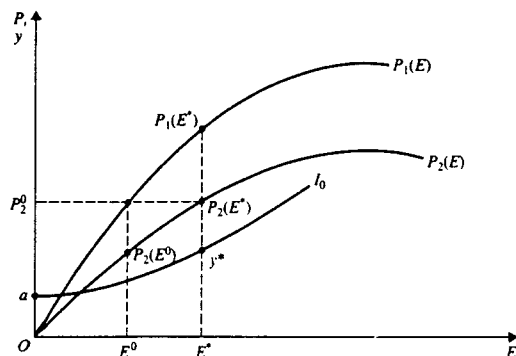


Fig. 13.3

if we assume:

1. O can never observe A 's choice of E directly: E is always A 's private information;
2. there is uncertainty: a number of possible values of P can result from a given choice of E .

These assumptions define a *hidden action* type of *principal-agent model* (see Chapter 22 for a general treatment). The implication of assumption (2) is that O cannot infer the value of E unambiguously from her observation of P . Fig. 13.3 illustrates this. For each E , suppose that the outcome P lies either on the curve $P_1(E)$ with probability π , or $P_2(E)$ with probability $1 - \pi$. Then if O observed the outcome P_2^0 , she would not be able to tell if it corresponded to an input of E^* and bad luck, or an input of E^0 and good luck. (Clearly, this problem would be compounded if there are more $P(E)$ curves with associated probabilities.)

Nevertheless, in this case O is able to write a contract enforcing choice of E^* or any E -value she desires. If A supplies any $E < E^*$, O is certain to observe a value of P other than $P_1(E^*)$ or $P_2(E^*)$, unless A chooses E^0 . If A chooses E^0 then with probability $(1 - \pi)$ O will observe $P_2(E^0) < P_2(E^*)$, and with probability π she will observe $P_1(E^0) = P_2(E^*)$. O then offers A the following contract. If she observes $P_1(E^*)$ or $P_2(E^*)$ she will pay y^* . However, if she observes $P_2(E^0)$ she pays \bar{y} , where \bar{y} satisfies

$$\pi v(y^*) + (1 - \pi)v(\bar{y}) - E^0 < u^0 \quad [\text{C.4}]$$

Finally, if she observes any other P , she pays $y < a$.

Faced with this contract A does best by supplying E^* . [C.4] ensures that he does worse by supplying the lower effort level E^0 and running the risk that an adverse state of the world will reveal his lack of effort (but see also Question 2, Exercise 13C).

O is able to enforce a desired effort level E^* in this case because some outcomes have zero probability if E^* is chosen and positive probability if some other E -value is chosen. Provided she can impose a large enough penalty (find a \bar{y} to satisfy [C.4]) she can always make $E \neq E^*$ not worthwhile for A . (Note that E^* may not actually be optimal for O in the above situation, but whatever E is optimal can be enforced in the same way.)

To formulate an agency problem which does not permit this type of solution, we proceed by making the probabilities of the outcomes, rather than the values of the outcomes themselves, functions of A 's effort level E . Suppose there are just two possible gross profit outcomes, P_1 and P_2 , with $P_1 > P_2$, and two possible effort levels A can choose, E_h and E_ℓ . We assume $\pi_h > \pi_\ell$, so that higher effort makes a higher profit outcome more likely. We assume that O seeks to maximize the expected value of profit, net of any payment she makes to A :

$$\bar{P}_i = \pi_i(P_1 - y_1) + (1 - \pi_i)(P_2 - y_2) \quad i = h, \ell \quad [\text{C.5}]$$

where π_i corresponds to A 's choice of E_i . Here we permit the payment to A to depend on the profit outcome, if necessary.

To induce A to accept the contract, O must ensure that A 's reservation expected utility constraint

$$\pi_i v(y_1) + (1 - \pi_i)v(y_2) - E_i \geq \bar{u}^0 \quad i = h, \ell \quad [\text{C.6}]$$

is satisfied for each choice of E_i . \bar{u}^0 is interpreted as the expected utility A would receive in his next best employment.

We begin by assuming that O can observe A 's choice of E_i . Then, just as before, she can write a contract forcing A to choose whichever E_i is better for her, and it remains only to find the payments O should make. Suppose O wants A to choose E_h . Then she finds y_1 and y_2 by solving

$$\max \bar{P}_h \quad \text{s.t.} \quad \pi_h v(y_1) + (1 - \pi_h)v(y_2) - E_h \geq \bar{u}^0 \quad [\text{C.7}]$$

with first-order conditions

$$-\pi_h + \lambda \pi_h v'(y_1^*) = 0 \quad [\text{C.8}]$$

$$-(1 - \pi_h) + \lambda (1 - \pi_h) v'(y_2^*) = 0 \quad [\text{C.9}]$$

$$\pi_h v(y_1^*) + (1 - \pi_h)v(y_2^*) - E_h = \bar{u}^0 \quad [\text{C.10}]$$

[C.8] and [C.9] yield

$$v'(y_1^*) = v'(y_2^*) = 1/\lambda \quad [\text{C.11}]$$

which in turn implies that $y_1^* = y_2^*$. Note also that $\lambda > 0$, so A 's reservation constraint must bind. Thus O holds A down to his minimum utility level, and compensates him for supplying E_h by paying him $y^* = y_1^* = y_2^*$. The result, that A receives a certain payment independently of the value of the gross profit outcome P_i , follows from the assumption that O is risk-neutral – she maximizes the expected value of her net profit (again see Chapter 19 for further discussion). Thus in this full information case A carries none of the risk associated with the enterprise.

We could now solve for the optimal payments when O wants A to choose the lower effort level E_ℓ , in exactly the same way. By comparing O 's value of \bar{P} at each of the two solutions, we could then find which E -level is better overall for O . However, we leave this as an exercise and assume that O would always want A to choose E_h .

Now suppose O cannot observe E . Then, since she cannot infer from the occurrence of a P -value which E has been chosen by A , she cannot write a (legally enforceable) contract forcing A to choose E_h . If she offers A the above payment, y^* , A can obviously make

himself better off by choosing E_L rather than E_h : his utility is going to be $v(y^*) - E_L$ for sure, and, from [C.10], $E_L < E_h$ must imply $v(y^*) - E_L > \bar{u}^0$. O can predict *ex ante* that under this contract A will choose E_L , but she cannot prove *ex post* that he has done so.

It follows that if O wants A to choose E_h , she must offer him a contract which gives him an incentive to do so – the contract must be *incentive compatible* with choice of E_h . We still wish to find y_1 and y_2 that maximize \bar{P}_h and still satisfy A 's reservation constraint [C.6] for $i = h$. However, to ensure it is in A 's interest actually to choose E_h we need to add to the problem in [C.7] the *incentive compatibility constraint*

$$\pi_h v(y_1) + (1 - \pi_h)v(y_2) - E_h \geq \pi_L v(y_1) + (1 - \pi_L)v(y_2) - E_L \quad [\text{C.12}]$$

This says that a pair of payments (y_1, y_2) must be chosen in such a way that A will prefer to choose E_h . (Note: to give ourselves a closed feasible set we assume that if A is indifferent between E_h and E_L for given (y_1, y_2) , he chooses E_h .)

Before solving this problem formally, note that [C.12] implies we *cannot* have $y_1 = y_2$, since $E_L < E_h$. In other words A is now going to have to carry some risk – his payment will vary with the gross profit outcome P . Adding [C.12] as a constraint to [C.7] yields

$$\frac{1}{v'(\hat{y}_1)} = \lambda + \mu \frac{(\pi_h - \pi_L)}{\pi_h} \quad [\text{C.13}]$$

$$\frac{1}{v'(\hat{y}_2)} = \lambda - \mu \frac{(\pi_h - \pi_L)}{(1 - \pi_h)} \quad [\text{C.14}]$$

$$\pi_h v(\hat{y}_1) + (1 - \pi_h)v(\hat{y}_2) - E_h = \bar{u}^0 \quad [\text{C.15}]$$

$$\pi_h v(\hat{y}_1) + (1 - \pi_h)v(\hat{y}_2) - E_h = \pi_L v(\hat{y}_1) + (1 - \pi_L)v(\hat{y}_2) - E_L \quad [\text{C.16}]$$

where \hat{y}_1, \hat{y}_2 are the optimal payments. It can be shown (see Question 5, Exercise 13C) that we must have $\lambda > 0$, $\mu > 0$, and so both constraints bind. Consider first [C.13] and [C.14]. Since $\pi_h > \pi_L$ (higher E increases the probability of the better outcome P_1), we must have

$$\frac{1}{v'(\hat{y}_1)} > \lambda > \frac{1}{v'(\hat{y}_2)} \quad [\text{C.17}]$$

or, since $v' > 0$,

$$v'(\hat{y}_1) < v'(\hat{y}_2) \quad [\text{C.18}]$$

Then, given $v'' < 0$ (diminishing marginal utility of income) we must have $\hat{y}_1 > \hat{y}_2$. In the light of [C.15] therefore

$$\hat{y}_1 > y^* > \hat{y}_2 \quad [\text{C.19}]$$

Thus, compared to the situation in which O can observe E , the payment to A now varies with the gross profit outcome: A receives less under P_2 , and more under P_1 , than in the full information contract. The reason for introducing this 'tilt' in the payment schedule is to give A an incentive to choose E_h rather than E_L : by doing so A increases the probability that he will receive the higher payment.

We obtain further insight into this by rearranging [C.16] to obtain

$$(\pi_h - \pi_L)[v(\hat{y}_1) - v(\hat{y}_2)] = E_h - E_L \quad [\text{C.20}]$$

The right-hand side is the cost to A of increasing his effort level from E_L to E_h . The left-hand side is the benefit: the probability of obtaining the higher outcome increases from π_L to π_h , and his utility increases from $v(\hat{y}_2)$ to $v(\hat{y}_1)$. The pair (\hat{y}_1, \hat{y}_2) must then be chosen to satisfy

$$v(\hat{y}_1) - v(\hat{y}_2) = (E_h - E_L)/[\pi_h - \pi_L] \quad [\text{C.21}]$$

where the right-hand side of this equation is determined by exogenous parameters of the problem. That is, \hat{y}_1 and \hat{y}_2 must be chosen to give A a sufficiently large expected utility increase to compensate for the increase in effort level. Note, in fact, that [C.21] and the reservation constraint [C.15] are entirely sufficient to determine the values of the two unknowns \hat{y}_1, \hat{y}_2 . It is unnecessary (in this case) to use [C.13] and [C.14]. (See Question 6, Exercise 13C.)

We can illustrate in Fig. 13.4 the two types of solution which arise depending on whether O can or cannot observe A 's effort level. Fig. 13.4 is an Edgeworth Box (see Chapter 16, section E and Chapter 17 for a fuller discussion of this type of diagram). The horizontal side of the box has length P_1 and the vertical side has length P_2 . The payment made by O to A if the profit outcome is P_1 is y_1 and is measured rightward from the origin O_A . Similarly, payment to A when the outcome is P_2 is measured up the vertical axis. Thus any point in the box measured from O_A shows the contract payments (y_1, y_2) to A and, measured from O_O , the net profits $(P_1 - y_1, P_2 - y_2)$ of the owner. The 45° line gives the points where $y_1 = y_2$ and A has a certain income: the payment he receives is the same whatever the profit outcome.

Indifference curves of A in Fig. 13.4, such as I_h^0 and I_L^0 , show combinations of y_1 and y_2 which give A the same expected utility for a given level of effort and so satisfy the equation

$$\pi_i v(y_1) + (1 - \pi_i)v(y_2) - E_i = \text{constant} \quad [\text{C.22}]$$

Applying the implicit function rule to [C.22] gives the slope of the indifference curves in y_1, y_2 space as

$$dy_2/dy_1 = -\pi_i v'(y_1)/(1 - \pi_i)v'(y_2) \quad [\text{C.23}]$$

Note that the slope of the indifference curves depends on the probability of the high profit outcome. Since high effort implies a greater probability of the large profit P_1 ($\pi_h > \pi_L$), the indifference curves of A are steeper at any point if he chooses a high effort level than if he chooses a low effort level. (Compare I_h^0 and I_L^0 at α or I_h^0 and I_L^0 at β .) Intuitively, he will require a smaller increase in y_1 to compensate him for a given reduction in y_2 the greater the probability that he attaches to getting y_1 .

O cares only about her expected income from the firm and so her indifference curves in the box show combinations of y_1 and y_2 which satisfy

$$\pi_i(P_1 - y_1) + (1 - \pi_i)(P_2 - y_2) = \text{constant} \quad [\text{C.24}]$$

The slope of her indifference curves is

$$dy_2/dy_1 = -\pi_i/(1 - \pi_i) \quad [\text{C.25}]$$

At the 45° line, incomes and therefore marginal utilities are equal, and so, from [C.23], the slope of A 's indifference curve is also given by [C.25]. Thus the straight line \bar{P}^* tangent

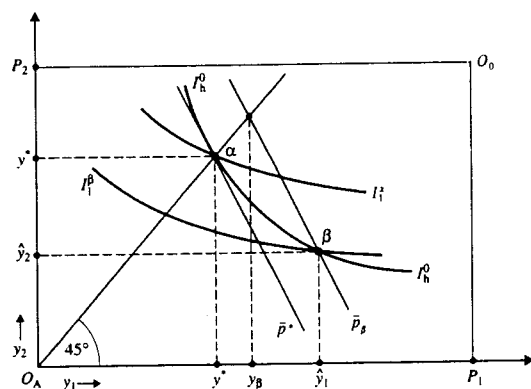


Fig. 13.4

to I_h^0 at α on the 45° line has slope

$$dy_2/dy_1 = -\pi_h/(1 - \pi_h) \quad [C.26]$$

and is an indifference curve for O given that A has chosen E_h . Clearly O prefers to be on indifference curves further from her origin O_0 since she then has a greater expected income. Note that along O 's indifference curves her expected income is constant, and so expected payments to A are also constant along her indifference curves.

When the owner can observe A 's effort the optimal contract is α . The indifference curve I_h^0 shows combinations of y_1 and y_2 which give A his reservation expected utility if he supplies high effort E_h :

$$\pi_h v(y_1) + (1 - \pi_h) v(y_2) - E_h = u^0$$

The contract α where I_h^0 crosses the 45° line therefore satisfies the condition [C.10] that A should get his reservation expected utility when he supplies high effort, and the condition [C.11] that $y_1 = y_2 = y^*$, so that he bears none of the risk.

Now consider the case in which O cannot observe A 's effort level and therefore must now provide an incentive compatible contract, which makes A choose E_h rather than E_l , even though his choice is not verifiable. If the contract is α then A will choose E_l rather than E_h since he gets the same certain income and his effort level is smaller. When the contract is α and A chooses E_l he is on the flatter indifference curve I_l^2 . To induce him to supply high effort the contract must give him a larger payment if there is a larger profit. Since we have seen that the reservation expected utility constraint will also bind, the optimal contract under asymmetric information must be on I_h^0 and below the 45° line so that $y_1 > y_2$. At contracts further to the right below the 45° line along I_h^0 , the expected utility that A gets from supplying low effort gets smaller. For example, compare the low effort indifference curves I_l^0 and I_l^1 – A clearly has a smaller expected utility, given that he chooses E_l , at β than at α because he is on a lower indifference curve. However, he is indifferent between the contracts α and β if he chooses E_h . By moving the contract down

I_h^0 , the owner will eventually find a contract where A 's expected utility from choosing E_l has fallen sufficiently that A is indifferent between E_h and E_l .

We assume that β is in fact the optimal contract under asymmetric information. The precise location of β will depend on the cost of additional effort ($E_h - E_l$) to A , and the slopes of his indifference curve, which in turn depend on the probabilities π_h and π_l and his marginal utility of income v' . However the diagram can be used to bring out one important implication: his possession of private information about his choice of E makes A no better off and O strictly worse off, than in the case in which O can observe E . Thus there is an *agency cost*, or cost of asymmetric information, to the less well informed party O which confers no benefit on the better informed party A .

The agency cost is the difference in O 's expected income from her optimal full information contract α and her optimal asymmetrical information contract β :

$$\begin{aligned} & \pi_h(P_1 - y^*) + (1 - \pi_h)(P_2 - y^*) - \pi_h(P_1 - \hat{y}_1) - (1 - \pi_h)(P_2 - \hat{y}_2) \\ & = \pi_h \hat{y}_1 + (1 - \pi_h) \hat{y}_2 - y^* \end{aligned} \quad [C.27]$$

In order to induce A to choose E_h when she cannot observe E , O must offer A a risky contract which gives him greater payment when profit is high. But, since A dislikes risk, the contract must also increase his expected income to compensate for the increased risk and to ensure that he gets his reservation expected utility. The agency cost to O is the increase in A 's expected income necessary to induce him to accept the incentive compatible contract β . This is [C.27]. In Fig. 13.4, the agency cost is measured by the distance between O 's indifference curves \bar{P}^* and P_β at the 45° line: $y_\beta - y^*$.

What does the analysis of this section tell us about the implications of the separation of ownership from control? First, the nature of the owner's information is crucial. For there to be a potential problem, O must be unable to observe or infer A 's choice of decision variable – there must be asymmetry of information. However, such asymmetry of information is not sufficient for agency costs to arise. If some outcomes have zero probability of being observed when A does what he should do, and positive probability of being observed when he does not, then provided that O can impose sufficiently large penalties, she can always ensure that A acts in her best interests. A genuine agency problem arises when there is no chance that O will have clear evidence that A has not acted in her best interest. In this case, she should offer A a contract which will provide him with an incentive to make his choices conform more closely to O 's interests. In the model of this section, the contract took an intuitively appealing form: the payment to A increases with the profit outcome of the business. When there is a separation of ownership from control, incentive contracts can be constructed to minimize the cost of asymmetric information but will not eliminate it entirely.

Exercise 13C

- Suppose that the optimal contract must maximize A 's expected utility subject to a constraint on O 's expected net profit. Derive the results for this case and compare them to those given in this section.

2. In equation [C.4], we define the payment \bar{y} as that which will induce A to choose $E = E^*$ rather than $E = E^0$. Will it always be possible to find such a \bar{y} ? If not, what kind of contract must O devise?
3. Solve the problem in [C.7] for the case in which O wants A to choose E_1 rather than E_2 . Is there in this case a problem for O when she cannot observe A 's choice of E ?
4. Suppose that instead of maximizing expected net profit, O seeks to maximize expected utility

$$\bar{U} = \pi U(P_1 - y_1) + (1 - \pi)U(P_2 - y_2)$$

where U is a cardinal utility function with $U' > 0$, $U'' < 0$. Show that in general A will not receive a payment that is independent of P even when O can observe choice of E .

5. Show that in conditions [C.13]–[C.16], we must have $\lambda > 0$, $\mu > 0$, i.e. [C.6] and [C.12] cannot be satisfied as inequalities.
6. *Comparative statics of contracts.* Use the fact that O 's optimal contract under asymmetric information is completely characterized by [C.15] and [C.16], to investigate the way in which changes in model parameters affect the optimal contract. In particular: what factors tend to increase agency costs?

D. Labour managed firms

In the standard model of the capitalist firm in earlier chapters workers are paid a fixed market-determined wage in exchange for their labour. The surplus of revenue over all payments to input suppliers accrues to the owners of the firm. However, there are many firms in which the workers own the firm, in the sense that they are rewarded for supplying labour by a share in the surplus of revenue over payments to all the non-labour inputs. Examples of such labour cooperatives or labour managed firms (LMF) include kibbutzim in Israel, Soviet cooperative farms, Basque industrial firms and partnerships of professionals such as lawyers. In this section we analyse the behaviour of such firms, and compare them with those of the standard capitalist firm, by outlining a model of the LMF first formulated by Ward (1958) and Vanek (1970).

We consider the LMF's short-run employment decision, with capital K fixed and employment, N , variable. The firm's production function is $q = f(N, \bar{K})$, with $f_N > 0$, $f_{NN} < 0$. Assume that it sells into a perfectly competitive market with given price p (it is easy to extend the model to the case of a LMF with monopoly power). Let F denote the fixed cost payable for the firm's capital stock \bar{K} . It is assumed that each worker receives an income y given by

$$y = (pq - F)/N \quad [\text{D.1}]$$

i.e. since the workers own the firm they share its profits. The second central assumption is that the firm seeks to maximize this income per worker. So, it solves

$$\max_N y = [pf(N, \bar{K}) - F]/N \quad [\text{D.2}]$$

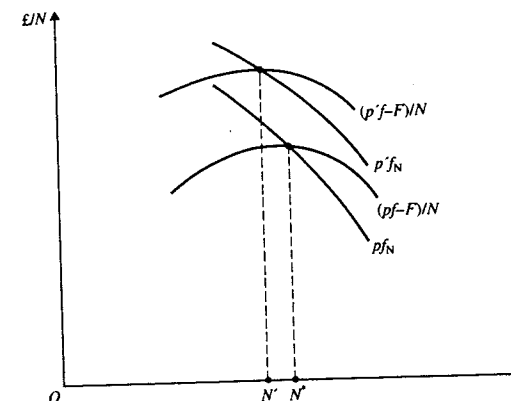


Fig. 13.5

giving the first-order condition

$$pf_N(N^*, \bar{K}) = [pf(N^*, \bar{K}) - F]/N^* = y^* \quad [\text{D.3}]$$

to determine optimal employment N^* . The firm sets employment at a level that equates the marginal value product of labour to income per worker. Employment will be expanded as long as an additional worker adds more to revenue than she is paid. Unlike the capitalist firm each worker is paid a profit share, rather than an externally determined market wage.

The comparative statics effect of a change in the market price p is startling. If we differentiate through [D.3] totally and rearrange we obtain

$$\frac{\partial N}{\partial p} = \frac{-(f_N - f/N)}{f_{NN}} < 0 \quad [\text{D.4}]$$

since $f_{NN} < 0$ implies $f_N < f/N$. An increase in the market price reduces the firm's employment level, and hence its output! The reason is illustrated in Fig. 13.5. When the market price rises from p to p' , the marginal value product of labour curve, which is the labour demand curve of a capitalist firm, shifts up, but so also does the curve of income per worker. Because of the presence of the fixed cost, the latter curve shifts upward by more than the former, and the result is that the equilibrium employment level, which, as [D.3] shows, is at the intersection of the two curves, falls. The cost of the marginal worker rises relative to the contribution to revenue she makes, and so a firm seeking to maximize income per worker would fire her!

Since this result follows directly from the formulation of the problem in [D.2], it is not surprising that criticism of the model has focused on the Ward–Vanek formulation of the firm's objective function. Suppose we were to formulate the objective of a capitalist firm as maximizing profit per unit of capital, which on the face of it does not seem at all unreasonable. Then the firm's demand for capital would have the same perverse characteristics as the LMF's demand for labour in the Ward–Vanek model. Why is the formulation of the maximand as the absolute amount of profit appropriate in one case,

and profit per unit of input inappropriate in the other? The resolution of this point by James Meade gives an interesting insight into the nature of the capitalist firm. The crucial issue is *discrimination*. The LMF in the Ward–Vanek model is a *non-discriminating* firm in the sense that a new worker receives the same income or profit share as existing workers. On the other hand, a capitalist firm *discriminates* among owners in the return it pays on capital: suppliers of new capital may well receive a lower rate of return than existing owners, and this creates an important difference in decisions on input levels.

To see this, consider the following simple example (based on Meade (1986)). The 10 original owners of a firm each put up £100 to provide an initial capital stock costing £1000 and yielding £2000 in profit. They have the opportunity to increase the profit of the company by £1600 if they install extra capital costing £1000. They turn to the stock market. The stock market rate of return is 20 per cent. That is, a firm with a profit of £3600 would be valued at $£3600/1.2 = £3000$. It follows that they can create 30 shares in the company, each worth £100, and sell 10 of them on the market to raise the required capital, dividing the remainder among themselves. The new shareholders will receive £1200 of the profit of the enlarged firm, to earn the market rate of return of 20 per cent. Each original owner is receiving £240 on his original investment of £100, a rate of return of 140 per cent. Clearly, it will pay the initial owners of the firm to expand capital, financed by issuing new shares, as long as the rate of return on the investment exceeds the market rate of return. In that case it makes sense to formulate the firm's maximand as the *absolute* difference between profit, discounted at the market rate of return, and the cost of the investment (this is simply the net present value of investment, extensively discussed in Chapter 15).

Suppose instead that there is a rule requiring *all* shareholders, old and new, to receive the same rate of return – the firm is non-discriminating in Meade's terminology. This implies that the £3600 profit of the enlarged firm would have to be divided equally between the suppliers of the initial £1000, and the suppliers of the next £1000: each £1 of capital subscribed now earns the same rate of return of 80 per cent. The initial shareholders now receive £1800 on this initial investment, which is less than the £2000 they receive if they do not bring in new shareholders. Thus they will not do so. This case corresponds to the Ward–Vanek LMF model.

This discussion suggests that the non-discriminating nature of the LMF is an important element in explaining its behaviour. For example, if new workers were hired at a market wage rate, rather than at a profit share equal to that of existing workers, then it is easy to show that employment will equate the marginal value product of labour to the wage rate and the perverse effects of changes in the output price disappear.

Other critics of the Ward–Vanek model focus on the narrow specification of the objective function which neglects social and ideological aspects of labour cooperatives. The LMF would generally be thought to have wider goals, and in particular a more explicit concern with employment as such, than can be captured by maximization of the income of the representative worker. For example, labour cooperatives often originate in an attempt to maintain employment in a firm that has gone bankrupt under conventional ownership.

To bring out the implications of such objectives suppose that there is a given population of G workers available to be employed in a LMF. If one of these workers is unemployed, she receives an income of b (say, unemployment benefit). The LMF has a *social* objective, which is the maximization of the total income accruing to all G workers. As in the Ward–Vanek model, the income of an employed worker is $y = [pf(N, \bar{K}) - F]/N$. The

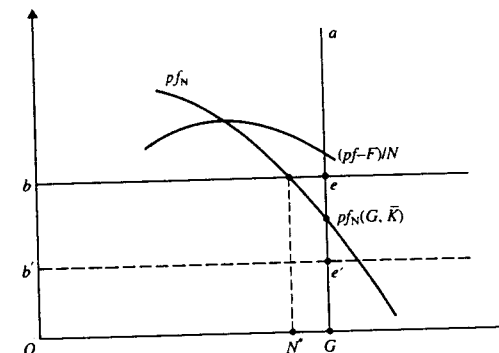


Fig. 13.6

firm's objective is now

$$\max_N Ny + (G - N)b \quad [D.5]$$

yielding the first-order condition

$$(y^* - b) + (pf_N - y^*) = 0 \quad [D.6]$$

$$\Rightarrow pf_N(N^*, \bar{K}) = b \quad \text{for } N^* < G \quad [D.7]$$

As long as an extra worker adds more to the firm's revenue than b , total income is increased by employing her. Moreover, differentiating through [D.7] now gives

$$\partial N^*/\partial p = -f_N/f_{NN} > 0 \quad [D.8]$$

and an increase in the market price increases employment in the firm. Fig. 13.6 illustrates this model (and shows also a solution where, for a low $b = b'$, $N^* = G$ and so $pf_N > b$, in which case $\partial N^*/\partial p = 0$).

Since workers will supply labour to the LMF, rather than remain unemployed, if they receive an income of at least b from the firm, we can regard the curve bea as a supply curve of labour. A conventional capitalist firm would also presumably be able to hire up to G workers by paying a wage of b and it would maximize profit by employing workers up to the point where $pf_N = b$. Thus the LMF with the social objective [D.5] employs the same number of workers as a profit maximizing firm. However, if the supply curve was $b'e'a$ the workers would be better off with a LMF than working for a capitalist firm. With the supply curve $b'e'a$ both types of firm would use the same number of workers G . The LMF would pay them each an income of $pf_N(G, \bar{K})$. But since G workers would be forthcoming for b' , the capitalist firm would only need to pay each of them $b' < pf_N(G, \bar{K})$. In this case the difference in objectives does make a difference because the LMF with the social objective does not attempt to exploit its monopsony power to keep the wage down to b' . Thus we have another example of the general lesson of this chapter: whether a firm will behave like a conventional profit maximizer will depend both on its objectives and on its environment.

Exercise 13D

1. Derive the results for the Ward–Vanek model in the case in which labour is the only input, so that $\bar{K} = F = 0$.
2. Suppose there are N_i ‘inside workers’ who control the decisions of the firm. They share the firm’s profit equally among themselves. All other workers are hired at the competitive market wage rate. Thus we have a ‘discriminating LMF’. Analyse its employment choice and the effect on this of changes in the output price.
3. The Ward–Vanek model could be interpreted as maximizing the utility of the ‘representative worker’, where this utility is linear in income and defined on no other variable. Suppose instead that each of the N workers may work ℓ hours, and that each possesses the identical quasi-linear utility function $v(y) - \ell$. Analyse the LMF’s choice of N and ℓ when it seeks to maximize the utility of the representative worker. What are the effects of a change in the output price in this case? (Hint: write the production function as $f(N\ell, \bar{K})$.)

E. Conclusions

There are many other non-standard models in which firms may not seek to maximize profit because their owners are interested in other aspects of the firm’s behaviour or are unable to fully control them to ensure profit maximization. The concluding set of exercises asks you to investigate the implications of the fact that the owners of firms are often consumers of their products. Chapter 13 of the workbook accompanying this text explores the sales maximization and expense preference models, proposed by W. J. Baumol and O. E. Williamson respectively, as two examples of what may happen when managers are imperfectly controlled by the firm’s owners. ‘Alternative’ models of the firm are of interest in their own right and also deepen our understanding of the conventional theory by focusing attention on what is required for the firm to pursue profit maximization.

Exercise 13E

1. *Consumers as owners.* A firm is the only producer of a good x , which sells at a price p and costs $c(x)$ to produce. The i th shareholder in the firm receives the share θ_i of its profit $\pi = px - c(x)$, but may also consume its product. Individual i has the utility function $u^i(x_i, y_i)$ where x_i is consumption of the firm’s product and y_i expenditure on all other goods and services (a composite commodity with a price of 1). Assume that all consumer-owners have preferences such that their income elasticity of demand for the firm’s product is zero. The budget constraint of the consumer-owner is $\bar{y}_i + \theta_i \pi \geq y_i + px_i$, where \bar{y}_i is income other than from the firm.
 - (a) Show that in their role as consumers, owner-consumers will act as if they each faced different prices $\hat{p}_i = p - \theta_i(p - c')$ where c' is the firm’s marginal cost.
 - (b) Show that the i th consumer-owner will wish the firm to set a price satisfying

$$\frac{p - c'}{p} = \frac{(\theta_i - \delta_i) 1}{\theta_i e}$$

where $\delta_i = x_i/x$ is the ratio of i ’s consumption to the total output of the firm

and e is the price elasticity of demand for the firm’s product. (Compare this with the standard monopoly price marginal cost margin in Chapter 11, section B.) Interpret this result – when will i wish the firm to maximize profit? – will i ever wish the price to be greater than the profit maximizing level, or less than marginal cost? – when will owners be able to agree on what price the firm should set?

2. *Taxpayer-consumers and public sector firms.* Suppose that the firm in the previous question is a public sector firm and that the public sector budget constraint is $G = \Pi + t\bar{y}$, where G is fixed government expenditure, $\bar{y} = \sum_i \bar{y}_i$ is the total income of the individuals in the economy and t is the proportional income tax rate. Thus increases in Π reduce the rate of income tax. Consumer-taxpayers have the budget constraint $\bar{y}_i(1 - t) \geq y_i + px_i$. What price will the i th consumer-taxpayer wish the public sector firm to set? (Hint: show that the previous analysis can be used with $\theta_i = \bar{y}_i/\bar{y}$.)
3. *Consumer cooperatives.* Consider the firm in question 1 being operated as a consumer cooperative in which all consumers are members. The profit is distributed to consumer-members by giving them a share equal to the ratio of their expenditure to the total revenue of the firm. Hence $\theta_i = px_i/px = x_i/x$ and consumers get a ‘dividend’ of $\pi x_i/x$. Show that consumer-members (a) act as if they were faced with a price equal to average cost $c(x)/x$ and (b) do not care what price the firm sets. (c) Why do consumers not wish to see $p = c'$ even though $\theta_i = \delta_i$?

References and further reading

The basic reference on the theory of the entrepreneurial firm is:

T. Scitovsky. ‘A note on profit maximization and its implications’, *Review of Economic Studies*, 11, 1943, 57–60.

For a fuller discussion of principal-agent models, see Chapter 22 and the references given there.

The references to work on the labour managed firm given in this chapter are:

J. E. Meade. *Alternative Systems of Business Organization and of Workers’ Remuneration*, Allen and Unwin, London, 1986.

J. Vanek. *The General Theory of Labor-Managed Market Economies*, Cornell University Press, Ithaca, NY, 1970.

B. Ward. ‘The Firm in Illyria: market syndicalism’, *American Economic Review*, 48, 1958, 566–89.

CHAPTER 14

Input markets and bargaining

In this chapter we examine aspects of markets for the inputs used in production by firms. Although the explicit focus will usually be on labour markets, much of the analysis will have a wider application. Section A addresses competitive input markets and considers the demand for inputs by profit maximizing firms who treat input prices as parameters. We do not examine the competitive supply of inputs since we have already covered this in Chapter 5, sections C and D (supply of labour by utility maximizing consumers) and Chapter 9 (supply by profit maximizing firms). Non-competitive input markets are dealt with in section B, where there is a single buyer of the input (monopsony), and in section C, where we consider unions as monopoly sellers of labour. Section D examines a bilateral monopoly, in which a monopoly union bargains with a single buyer of labour, and sets out the efficient bargain model. This is an example of the cooperative game approach to bargaining which attempts to predict what bargains will be made by laying down reasonable criteria which bargains must satisfy. Section E continues the cooperative game analysis of bargaining in a more general setting and shows how four reasonable axioms yield a unique bargaining outcome for a wide class of cooperative bargaining games.

Section F introduces the contrasting non-cooperative game approach which examines the process of bargaining and the parties' bargaining strategies to predict the bargain which will be made. Finally, in section G, we show how imperfect information can result in delay in reaching agreement or even in no agreement at all being made.

A. Demand for inputs

We will concentrate on the demand for inputs by a profit maximizing firm facing input prices which it regards as unalterable by its actions. There are assumed to be no adjustment costs involved in varying input levels or, in the terminology of Chapters 7, 8 and 9, the firm's problem is *long-run*: there are no constraints on the adjustment of its inputs. Derivation of the short-run demand for inputs is left to the exercises. To keep the analysis simple it is further assumed that the firm produces a single output y from two inputs z_1, z_2 subject to the constraint $y \leq f(z_1, z_2)$, where f is a production function. Since a profit

maximizing firm never produces where $y < f(z_1, z_2)$ (explain why) the production constraint can be treated as an equality: $y = f(z_1, z_2)$. The firm faces a demand curve for its output, $p = p(y)$. If $dp/dy = 0$ the demand curve is horizontal and the firm sells y in a competitive market. If $dp/dy < 0$ the demand curve is negatively sloped and the firm is a monopolist. The firm's total revenue is $R(y) = p(y)y$ and since the production constraint is an equality we can write $R(y) = R[f(z_1, z_2)]$. Since choice of z_1, z_2 determines costs and revenue the firm's output need not appear explicitly in its profit maximization problem

$$\max_{z_1, z_2} R[f(z_1, z_2)] - \sum_i p_i z_i \quad [\text{A.1}]$$

where p_i is the price of z_i . (Compare equation [A.3] in Chapter 9, which relates only to the competitive case.)

Assuming that both inputs are positive at the solution, necessary conditions for a maximum are

$$R'f_i - p_i = 0 \quad (i = 1, 2) \quad [\text{A.2}]$$

where $R' = dR/dy$ is marginal revenue and f_i is the marginal product of z_i in the production of y . [A.2] can be rewritten as

$$MR \cdot MP_i = p_i \quad (i = 1, 2) \quad [\text{A.3}]$$

The firm will adjust its input levels until the cost of an extra unit of input i , p_i , is equal to the extra revenue generated by the extra unit, $MR \cdot MP_i$. The increase in z_i increases y by MP_i (its marginal product) and a unit increase in output increases revenue by marginal revenue. $MR \cdot MP_i$ is usually called the *marginal revenue product* of z_i and written MRP_i . When the firm sells y in a competitive market

$$MR = \frac{dp}{dy} \cdot y + p = p$$

since dp/dy is zero. In this case the MRP_i is $p \cdot MP_i$ which is known as the *value of the marginal product* and written VMP_i . Given that dp/dy is non-positive we see that $VMP_i \geq MRP_i$.

Recalling from Chapter 8, section B that p_i/MP_i is marginal cost, if we divide both sides of [A.3] by MP_i we get

$$MR = MC = \frac{p_i}{MP_i} \quad (i = 1, 2) \quad [\text{A.4}]$$

and so we have the familiar conclusion that profit maximization requires that marginal revenue be equated to marginal cost.

Profit maximization also requires that the cost of any given output level be minimized. Dividing the profit maximizing condition [A.3] on input 1 by the profit maximizing condition on input 2 gives

$$\frac{MP_1}{MP_2} = \frac{p_1}{p_2}$$

which is just the requirement for cost minimization: the firm chooses an input combination where its isoquant is tangent to its isocost line (recall Chapter 7, section B).

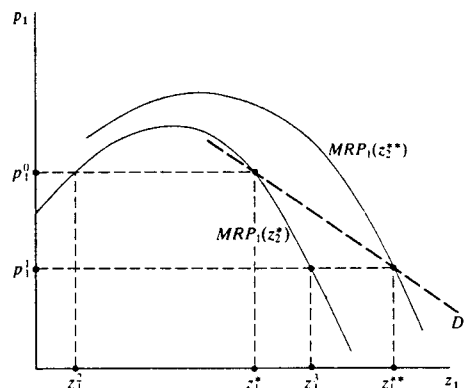


Fig. 14.1

From the equilibrium conditions [A.2] we see that the firm's demand for inputs will depend on the prices of the inputs and the parameters of the production and output demand functions. Let us consider how the demand for an input varies with its price. Denote the initial price of z_1 by p_1^0 . At this price, and given the price of z_2 , the firm chooses the initial optimal combination (z_1^*, z_2^*) . If z_2 is held constant at z_2^* , then $MRP_1 = R'[f(z_1, z_2^*)] \cdot f_1(z_1, z_2^*)$ varies only with z_1 in Fig. 14.1. This is the curve labelled $MRP_1(z_2^*)$, to indicate that its position depends on the pre-assigned level of z_2 . As z_1 varies with z_2 fixed, MRP_1 varies, first because more output is produced and this will reduce MR if the firm faces a negatively-sloped demand curve in its output market; and second because MP_1 varies with z_1 . Now over a range of values of z_1 , MP_1 may rise with z_1 (see Chapter 7, section C) and so it is possible that MRP_1 at first rises with z_1 (the increase in MP_1 offsetting any decrease in MR) and then falls (the MP_1 must eventually decline and so reinforce the nonpositive change in MR). The $MRP_1(z_2^*)$ curve in Fig. 14.1 reflects this possibility.

The firm chooses its profit maximizing level of z_1 where $p_1^0 = MRP_1$. But MRP_1 equals p_1^0 at both z_1^* and z_1^1 . At z_1^1 however the $MRP_1(z_2^*)$ curve cuts the p_1^0 line from below, indicating that an increase in z_1 above z_1^1 will lead to $MRP_1 > p_1$, i.e. an increase in z_1 will generate revenue in excess of its cost. Hence z_1^1 cannot be the optimum. At z_1^* on the other hand, an increase in z_1 will lead to $MRP_1 < p_1$ so that profit is reduced and a reduction in z_1 loses more revenue than cost (since then $MRP_1 > p_1$). Hence the profit maximizing level of z_1 must occur where MRP_1 is negatively sloped and cuts the p_1^0 line at z_1^* .

Now suppose that p_1 falls to p_1^1 , so that the price line cuts $MRP_1(z_2^*)$ at z_1^1 . Is this the new profit maximizing level of z_1 ? The answer is no, because a change in p_1 will also cause a change in the optimal z_2 , to z_2^{**} , so that the MRP_1 curve will shift to the right to $MRP_1(z_2^{**})$. The new optimal level for z_1 is z_1^{**} where $p_1^1 = MRP_1(z_2^{**})$. The demand curve for z_1 is therefore the negatively sloped dashed line D_1 : a fall in the price of an input leads to an increased demand for it by the firm.

If $MRP_1(z_2^{**})$ lies to the right of $MRP_1(z_2^*)$ the demand curve must be negatively sloped.

We have merely asserted rather than proved that this is the case. It is tedious to show that MRP_1 shifts right as p_1 falls and so we will support our conclusion that D_1 is negatively sloped by two more direct arguments. For a firm selling y on a competitive market we have already proved that input demand functions are negatively sloped (Chapter 9, section D) but the arguments used here also apply to the case of monopoly in the output market and have instructive similarities to the methods of Chapter 4, section B and Chapter 8, section B.

The first demonstration that the firm's demand curve for inputs is negatively sloped makes use of the properties of the maximum profit function. Since the optimal input demands depend on the input prices p_i

$$z_i^* = D_i(p_1, p_2) \quad (i = 1, 2) \quad [\text{A.5}]$$

so does the firm's maximum profit:

$$\pi_{\max} = R(f(z_1^*, z_2^*)) - \sum p_i z_i^* = \pi^*(p_1, p_2) \quad [\text{A.6}]$$

Differentiating $\pi^*(p_1, p_2)$ with respect to p_k gives

$$\frac{\partial \pi^*}{\partial p_k} = R' \left\{ f_1 \frac{\partial z_1^*}{\partial p_k} + f_2 \frac{\partial z_2^*}{\partial p_k} \right\} - \sum p_i \frac{\partial z_i^*}{\partial p_k} - z_k^* \quad [\text{A.7}]$$

But from [A.2], $R'f_i = p_i$ and so rearranging [A.7] gives

$$\frac{\partial \pi^*}{\partial p_k} = \sum_i (R'f_i - p_i) \frac{\partial z_i^*}{\partial p_k} - z_k^* = -z_k^* = -D_k(p_1, p_2) \quad [\text{A.8}]$$

This is yet another example of the Envelope Theorem (Chapter 2, section J) and shows that for all profit maximizing firms Hotelling's lemma holds in respect of input prices, whether the firms operate in a competitive or monopolized output market (recall Chapter 9, section D for the competitive firm). If we can also show that $\pi^*(p_1, p_2)$ is convex in input prices so that

$$\frac{\partial^2 \pi^*}{\partial p_k^2} = -\frac{\partial D_k}{\partial p_k} \geq 0 \quad [\text{A.9}]$$

then we will have completed our first demonstration that input demand curves are (weakly) negatively sloped, i.e. $\partial D_k / \partial p_k \leq 0$.

To establish the convexity of $\pi^*(p_1, p_2)$, consider three input price vectors p^0 , p^1 and $\bar{p} = tp^0 + (1-t)p^1$ ($0 \leq t \leq 1$) and the three corresponding profit maximizing input vectors z^0 , z^1 and \bar{z} . We have

$$\pi^*(p^0) = R(f(z^0)) - \sum p_i^0 z_i^0 \geq R(f(\bar{z})) - \sum p_i^0 \bar{z}_i \quad [\text{A.10}]$$

and

$$\pi^*(p^1) = R(f(z^1)) - \sum p_i^1 z_i^1 \geq R(f(\bar{z})) - \sum p_i^1 \bar{z}_i \quad [\text{A.11}]$$

Multiplying through [A.10] by t and [A.11] by $(1-t)$ and then adding the corresponding

sides of the two resulting inequalities gives

$$\begin{aligned} t\pi^*(p^0) + (1-t)\pi^*(p^1) &\geq t[R(f(\bar{z})) - \sum p_i^0 \bar{z}_i] + (1-t)[R(f(\bar{z})) - \sum p_i^1 \bar{z}_i] \\ &= R(f(\bar{z})) - \sum [tp_i^0 + (1-t)p_i^1] \bar{z}_i \\ &= R(f(\bar{z})) - \sum \bar{p}_i \bar{z}_i = \pi^*(\bar{p}) \end{aligned}$$

so that π^* is indeed convex in input prices and [A.9] holds.

Substitution and output effects

Our second demonstration of the negative slope of the firm's input demand curves requires a more detailed consideration of the effect of a change in the price of an input on the firm's behaviour. The firm's demand for z_1 changes as p_1 changes, first because a *different input combination* will now minimize the cost of any given output and second because a *different output level* will now be optimal. We can call these two effects the *substitution* and *output effects* of a change in p_1 . In section 8B we showed that the substitution effect of an input price fall always leads to a rise in the use of the input whose price had fallen. We now use techniques similar to those used in deriving the Slutsky equation of Chapter 5, section B, to decompose the total effect of a change in p_1 into the substitution and output effects.

From Chapter 8, section B we know that the *cost minimizing* z_1 depends on the input prices and the level of y :

$$\hat{z}_1 = h_1(p_1, p_2, y) \quad [\text{A.12}]$$

where \hat{z}_1 denotes the cost minimizing z_1 . From the firm's profit maximization problem [A.1] we know that the *profit maximizing* z_1 depends on p_1 and p_2 , as shown in [A.5], and that the profit maximizing output y^* will therefore also depend on p_1 and p_2 since choice of z_1, z_2 determines y :

$$y^* = f(z_1^*, z_2^*) = y^*(p_1, p_2) \quad [\text{A.13}]$$

If we set y in [A.12] equal to y^* in [A.13]:

$$y = y^*(p_1, p_2) \quad [\text{A.14}]$$

then, since profit maximization implies cost minimization, it must be true that

$$z_1^* = D_1(p_1, p_2) = \hat{z}_1 = h_1(p_1, p_2, y^*(p_1, p_2)) \quad [\text{A.15}]$$

Now let p_1 vary but ensure that y in $h_1(p_1, p_2, y)$ varies to maintain the equalities in [A.14] and [A.15]. Hence h_1 will vary, first because with y constant a new cost minimizing input combination is chosen and second because varying p_1 will change y^* and therefore y via [A.14]. Differentiating [A.15] with respect to p_1 gives

$$\frac{\partial D_1}{\partial p_1} = \frac{\partial h_1}{\partial p_1} + \frac{\partial h_1}{\partial y} \cdot \frac{\partial y}{\partial y^*} \cdot \frac{\partial y^*}{\partial p_1} \quad [\text{A.16}]$$

where the first term on the right-hand side of [A.16] shows how z_1 varies with p_1 when

y is constant and so is the substitution effect. The second term is the rate at which z_1 varies indirectly with p_1 because of the effect of changes in p_1 on the optimal output level. This is the output effect. From Chapter 8, section B we know that

$$\frac{\partial h_1}{\partial p_1} < 0$$

so let us consider the output effect. $\partial h_1 / \partial y$ is the rate at which z_1 varies with y along the cost minimizing expansion path and $\partial h_1 / \partial y$ may be positive (z_1 is normal) or negative (z_1 is an inferior input). From [A.14], $\partial y / \partial y^* = 1$. The last part of the second term is $\partial y^* / \partial p_1$: the rate at which the profit maximizing output varies with the price of input 1. Now, recalling equation [A.4] above, y^* is determined by the equality of marginal revenue with marginal cost. If a rise in p_1 shifts the marginal cost curve upward then output must fall and, conversely, if the marginal cost curve falls as p_1 rises output will rise. Hence $\partial y^* / \partial p_1$ is positive or negative as MC falls or rises with p_1 , i.e. as $\partial MC / \partial p_1$ is negative or positive. But from Chapter 8, section B, $\partial MC / \partial p_1$ is negative or positive as z_1 is inferior or normal. Hence

$$\frac{\partial h_1}{\partial y} \geq 0 \Leftrightarrow \frac{\partial MC}{\partial p_1} \geq 0 \Leftrightarrow \frac{\partial y^*}{\partial p_1} \leq 0$$

and therefore

$$\frac{\partial h_1}{\partial y} \cdot \frac{\partial y}{\partial y^*} \cdot \frac{\partial y^*}{\partial p_1} = \frac{\partial h_1}{\partial y} \cdot \frac{\partial y^*}{\partial p_1} < 0 \quad [\text{A.17}]$$

The output effect of a rise in p_1 always reduces the demand for z_1 , so reinforcing the substitution effect. We have therefore established by another route that

$$\frac{\partial D_1}{\partial p_1} = \frac{\partial h_1}{\partial p_1} + \frac{\partial h_1}{\partial y} \frac{\partial y^*}{\partial p_1} < 0 \quad [\text{A.18}]$$

i.e. the input demand curve is negatively sloped, irrespective of whether the firm sells its output in a monopolized or competitive market.

From [A.18] we see that the slope of the firm's input demand curve will depend on the magnitude of the substitution and output effects. The substitution effect will in turn depend on the curvature of the firm's isoquants and if the elasticity of substitution (Chapter 8, section B) is taken as the measure of curvature, the substitution effect is larger the larger is the elasticity of substitution. The output effect is the product of two terms and is larger the greater is the response of the cost minimizing level of z_1 to changes in output and the greater is the response of output to the change in input price. This latter influence will depend on how much marginal cost varies with the price of the input: the bigger the shift in the marginal cost curve the bigger the change in the profit maximizing output. If the firm is a monopolist in the output market the change in y will also depend on the slope of the marginal revenue curve: the steeper this is the smaller will be the change in y as the marginal cost curve shifts. (Draw a diagram to show this.)

The market input demand curve

The market demand curve for a consumer good is derived by horizontally summing the individual demand curves. If an input is used only by firms which are monopolists in their respective output markets and these are unrelated in demand then the input market demand curve can be derived in the same way by horizontal summation of the individual firms' demand curves (see Question 4, Exercise 14A). Apart from this somewhat unlikely case the input market demand curve is *not* the horizontal sum of individual firms' demand curves. The reason for this can be seen if we examine an input used only in production of one type of good which is sold on a competitive market by the many firms producing it. Consider Fig. 14.2, in which the curve ΣD_1^{j0} is the horizontal sum of the individual firms' demand curves for input 1 and, at the initial price p_1^0 , Σz_1^{j0} is demanded. Each individual demand curve is like the dashed line D_1 in Fig. 14.1, which shows how each firm's *ceteris paribus* demand varies with p_1 . It is assumed in drawing D_1 that the firm regards the price of output as unalterable by its actions, so that the D_1^{j0} curve of each firm is derived with the price of output held fixed. Hence the ΣD_1^{j0} curve is also based on the assumption that the price of output is constant. But when the input price p_1 falls to p_1^1 all firms' average and marginal cost curves alter. In the long run when the number of firms and the size of firms' plants can be varied the change in total output is determined by the change in the firms' average cost curves (see Chapter 10). Average cost curves shift down when the price of the input falls (Chapter 8, section B) and the long-run supply of the industry will increase. The price of output will therefore fall, shifting the MRP_1 , D_1^j and ΣD_1^j curves to the left. This is shown in Fig. 14.2, where ΣD_1^{j1} is the new horizontal sum of the new individual D_1^{j1} curves and the amount of input demanded at p_1^1 is Σz_1^{j1} . We see that the market input demand curve is D , which is steeper than the ΣD_1^j curves. (Compare the derivation of the market supply curves in Chapter 10.)

The market input demand curve is therefore determined by the demand conditions in the market for the output produced by the input, the change in firms' cost curves caused by the change in the input price and by the elasticity of substitution amongst the inputs.

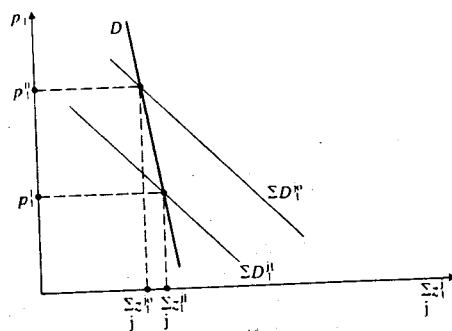


Fig. 14.2

Elasticity of market input demand

It is possible to derive a simple expression for the elasticity of the market input demand function in the case in which all firms have the same constant returns to scale production function. (In this case the size of the individual firm is indeterminate but this does not matter since we are only concerned with the input demand of all the firms together.) Since all production functions are identical and all firms face the same input prices, all firms have the same cost function $C(p_1, p_2, y^j)$, where y^j is the output of the j th firm. Because there are constant returns to scale the cost of producing y^j units is merely y^j times the cost of producing one unit. Letting $\bar{C}(p_1, p_2) = C(p_1, p_2, 1)$ be the cost function for producing one unit, the total industry cost for producing $y = \Sigma y^j$ is

$$\sum C(p_1, p_2, y^j) = \sum y^j \bar{C}(p_1, p_2) = y \bar{C}(p_1, p_2) \quad [\text{A.19}]$$

For industry equilibrium we require that (a) all firms are maximizing their profits, (b) every firm is just breaking even and (c) industry output, i.e. supply, equals demand. Now profit maximization implies cost minimization so that the j th firm's demand for z_1 is

$$z_1^j = C_1(p_1, p_2, y^j) = y^j \bar{C}_1(p_1, p_2) \quad [\text{A.20}]$$

where $\bar{C}_1 = \partial \bar{C} / \partial p_1$, and so the industry demand for z_1 is

$$z_1 = \sum z_1^j = y \bar{C}_1(p_1, p_2) \quad [\text{A.21}]$$

If all firms just break even then

$$p = \bar{C}(p_1, p_2) \quad [\text{A.22}]$$

where p is the price of the output produced by the industry. Since supply equals demand in the output market

$$y = D(p) \quad [\text{A.23}]$$

where $D(p)$ is the demand function for the industry's output. Substituting [A.22] and [A.23] into [A.21] gives the industry demand for z_1 as a function of the input prices only:

$$z_1 = D(\bar{C}(p_1, p_2)) \cdot \bar{C}_1(p_1, p_2) \quad [\text{A.24}]$$

Differentiation of [A.24] with respect to p_1 gives

$$\frac{\partial z_1}{\partial p_1} = D' \cdot \bar{C}_1 \cdot \bar{C}_1 + D \cdot \bar{C}_{11} \quad [\text{A.25}]$$

where $D' = dy/dp$ and $\bar{C}_{11} = \partial^2 \bar{C} / \partial p_1^2$. Multiplying [A.25] through by $-p_1/z_1$ and using [A.21] gives the elasticity of demand for z_1 with respect to its price

$$e_1 = \frac{-\partial z_1}{\partial p_1} \frac{p_1}{z_1} = \frac{-dy}{dp} \frac{p}{y} \frac{p_1 z_1}{p y} - \frac{y p_1}{z_1} \bar{C}_{11} = e \cdot s_1 - \frac{y p_1}{z_1} \bar{C}_{11} \quad [\text{A.26}]$$

where e is the price elasticity of demand in the product market and s_1 is the share of total cost spent on z_1 . The second term in the expression is less easy to interpret as it stands. However, recalling the discussion of the individual firm's demand and [A.18], \bar{C}_{11} is the substitution effect, showing how z_1 varies with p_1 with output constant. The second term therefore depends on the shape of the isoquant, i.e. the elasticity of substitution. In fact

we can show that the second term can be written as $s_2 \cdot \sigma$, where s_2 is the proportion of cost spent on the second input and σ is the elasticity of substitution. (See Question 1, Exercise 14A.) Hence the own price elasticity of demand for z_1 can be written as

$$e_1 = s_1 e + s_2 \sigma \quad [\text{A.27}]$$

The own price elasticity of demand for the input by the industry is a weighted average (since $s_1 + s_2 = 1$) of the price elasticity of demand product market and the elasticity of substitution. The output effect $s_1 e$ depends on how responsive cost and therefore the price of output is to changes in input prices (recall from Chapter 8 that s_1 is the elasticity of total and average cost and, in the constant returns case, marginal cost to p_1) and on how demand (equals output) varies with market price.

Exercise 14A

1. Show that the term $-(yp_1/z_1)\bar{C}_{1,1}$ in [A.26] can indeed be written as $s_2\sigma$. Generalize [A.27] to the case of the more than two inputs. (See Hicks (1963) and Diewert (1971), and Question 4, Exercise 7B).
2. The D curve in Fig. 14.2 is the long-run market demand curve for the input since it shows how demand varies when all inputs and the number of firms are freely variable. Construct the short-run market demand curve showing how demand varies with the price of the input when the other input is fixed and the number of firms does not alter. Would you expect this curve to be more or less elastic than that in Fig. 14.2?
3. Suppose that the fall in p_1 leads to a shift in the market demand curve for z_2 . Under what circumstances will this change the price of z_2 ? What effect will this have on the market demand curve for z_1 ?
4. Explain why, when the buyers of an input are monopolists in markets unrelated to each other in demand, the input market demand curve can be obtained by horizontal summation of the individual firms' demand curves.
5. *Substitutes and complements.* Input i is a substitute (complement) for input k if an increase in p_k increases (reduces) the firm's profit maximizing demand for input i . If i is a substitute for k does this imply that k must be a substitute for i ? (Compare the definition of Marshallian substitutes and complements for the utility maximizing consumer.)

B. Monopsony

Monopsony is defined as a market in which there is a single buyer of a commodity who confronts many sellers. Each of the sellers treats the market price of the good as a parameter and so there is a market supply curve for the good which is derived in the usual way from the supply curves of the individual suppliers. The single buyer of the good faces a market supply function relating total supply to the price he pays. This can be expressed (in the

inverse form) as:

$$p_1 = p_1(z_1) \quad (p'_1 > 0) \quad [\text{B.1}]$$

where [B.1] shows the price of the commodity which must be paid to generate a particular supply. Note that the buyer is assumed to face an upward-sloping supply curve; the price required is an increasing function of the amount supplied.

The market price of the monopsonized input is determined, given the supply function [B.1], by the buyer's demand for z_1 . We assume that the monopsonist is a profit maximizing firm, in which case the demand for z_1 , and hence its price, is determined by the firm's profit maximizing decision. In the two-input single-output case the firm's problem is

$$\max_{z_1, z_2} R[f(z_1, z_2)] - p_1(z_1)z_1 - p_2z_2 \quad [\text{B.2}]$$

This is very similar to problem [A.1] except that p_1 depends on z_1 because of [B.1]. Input 2 is assumed to be bought on a market in which the firm treats p_2 as a parameter. The firm's output may be sold in a competitive or a monopolized market: monopsony need not imply monopoly. The firm may, for example, be the only employer of labour in a particular area but be selling its output in a market where it competes with many other firms, and labour may be relatively immobile.

Necessary conditions for a maximum of [B.2] are (when both z_1 and z_2 are positive at the optimum):

$$R'_1 f_1 - (p_1 + p'_1 z_1) = 0 \quad [\text{B.3}]$$

$$R'_1 f_2 - p_2 = 0 \quad [\text{B.4}]$$

[B.4] is identical with [A.2], but [B.3] is not, because of the $p'_1 z_1$ term. The firm will adjust its use of an input up to the point at which the additional revenue from a unit of the input equals the extra cost incurred. When the price of the input is independent of the number of units bought the cost of an extra unit is its price. But when the firm faces an upward sloping supply curve for the input it must pay a higher price for *all* units bought to ensure supply for an extra unit. This means that the cost of an extra unit of z_1 is the price paid for that unit *plus* the increased cost of the units already bought, which is the rise in p_1 times the amounts of z_1 bought: $p'_1 z_1$. Hence writing MRP_i for the marginal revenue product of input i and MBC_i for the marginal cost of z_1 to the buyer (marginal buyer cost) the firm maximizes profits by setting

$$MRP_1 = MBC_1 > p_1 \quad [\text{B.5}]$$

$$MRP_2 = MBC_2 = p_2 \quad [\text{B.6}]$$

This equilibrium is illustrated for the monopsonized input in Fig. 14.3. S_1 is the supply curve of z_1 and MBC_1 plots the marginal buyer cost ($p_1 + p'_1 z_1$) of the single buyer. $MRP_1(z_1^*)$ is the marginal revenue product curve for the input given the optimal level of z_2 . The firm maximizes profit with respect to z_1 by equating MRP_1 to MBC_1 at z_1^* . To generate this supply of z_1 the firm will set the monopsony price $p_1^* = p_1(z_1^*)$.

The analysis of the single buyer confronting many competitive sellers is rather similar to the analysis in Chapter 11 of the single seller confronting many competing buyers. In each case the firm realizes that it faces a curve relating price to quantity which summarizes

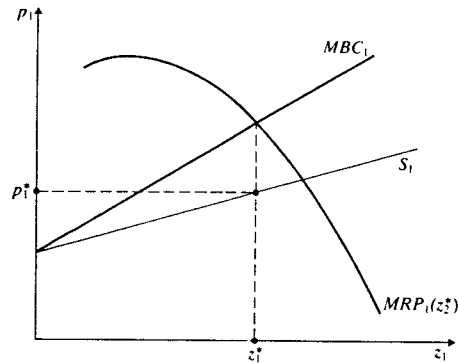


Fig. 14.3

the response of the competitive side of the market and the firm sets the quantity or price in the light of this interdependence of price and quantity. In each case the market price overstates the *marginal* profit contribution of the quantity and in each case this overstatement depends on the responsiveness of quantity to changes in price. Under monopoly the firm equates $MR = p[1 + (1/e)]$ to the marginal cost of output, and the less elastic is demand the greater is the difference between price and marginal cost. [B.5] can be rewritten in a similar way. Defining the elasticity of supply of z_1 with respect to price as

$$e_1^s = \frac{dz_1}{dp_1} \cdot \frac{p_1}{z_1} \quad [\text{B.7}]$$

we see that

$$MBC_1 = p_1 + \frac{dp_1}{dz_1} \cdot z_1 = p_1 \left(1 + \frac{1}{e_1^s} \right) \quad [\text{B.8}]$$

and so [B.5] becomes

$$MRP_1 = p_1 \left(1 + \frac{1}{e_1^d} \right) \quad [\text{B.9}]$$

The less elastic is supply with respect to price the greater will be the difference between MRP_1 and the price of the input. In other words, the less responsive to price the input supply is, the greater the excess of the value of the marginal unit of the input over the price it receives. This could be regarded as a measure of the degree of 'monopsonistic exploitation'.

An indifference curve analysis of monopsony

Further insight into the monopsony outcome can be gained by recasting the analysis in terms of indifference curves. Suppose for simplicity that the monopsonized input is the

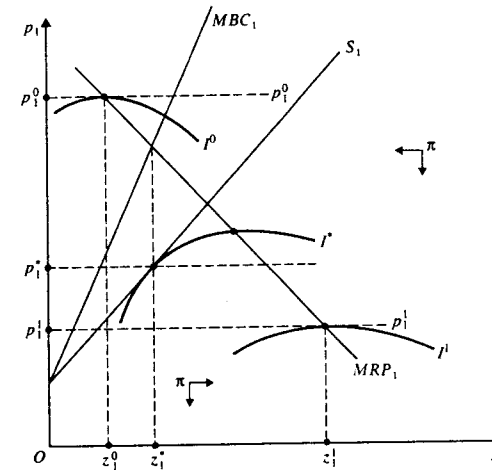


Fig. 14.4

sole input in the production of the monopsonist's output: $y = f(z_1)$. (Allowing for several inputs makes no essential difference to the analysis. See Question 5, Exercise 14B.) The revenue from the sale of y is $R(f(z_1))$ and the MRP_1 curve in Fig. 14.4 plots the marginal revenue product of the input: $Rf_1 \equiv MR \cdot MP_1$. The competitive supply curve is S_1 and MBC_1 is the corresponding marginal buyer cost curve.

The firm's objective function is the profit function

$$\pi = R(f(z_1)) - p_1 z_1 = \pi(z_1, p_1) \quad [\text{B.10}]$$

The effect on profit of a marginal increase in z_1 with p_1 held constant is $\pi_{z_1} = Rf_1 - p_1$. Holding the input constant, an increase in the input price reduces profit at the rate $\pi_{p_1} = -z_1$. Hence the firm's indifference curves in (p_1, z_1) space have slope

$$\left. \frac{dp_1}{dz_1} \right|_{d\pi=0} = -\frac{\pi_{z_1}}{\pi_{p_1}} = \frac{Rf_1 - p_1}{z_1} \quad [\text{B.11}]$$

Since π is decreasing in p_1 , lower indifference curves in Fig. 14.4 correspond to higher profit: the firm is better off on I^1 than on I^0 . The firm will maximize profit by choosing the (p_1, z_1) combination which gets it onto the lowest feasible indifference curve. What (p_1, z_1) combinations are feasible depend on the input market conditions which the firm faces.

If the firm was a competitive buyer of the input it would treat p_1 as a parameter: it would be constrained to choose a (p_1, z_1) combination on the horizontal line with height equal to the given p_1 . For example, if $p_1 = p_1^0$ it would maximize profit by choosing an input level of z_1^0 where its indifference curve I^0 is tangent to the horizontal line p_1^0 . Hence the slope of its indifference curve at the profit maximizing (p_1, z_1) combination would be zero and so $Rf_1(z_1^0) = p_1^0$. Similarly, if the firm acted as if the input price was a parameter

equal to p_1^1 , it would choose the profit maximizing input level z_1^1 , where I^1 is tangent to the horizontal line $p_1^1 p_1^1$ and its marginal revenue product is equal to the input price. The firm's demand curve for the input is the locus of such points of tangency, i.e. its MRP_1 curve. Notice that to the left of the MRP_1 curve the firm's indifference curves are positively sloped and to the right of it they are negatively sloped. This follows from the fact that to the left of MRP_1 increases in z_1 at given p_1 increase profit and to the right of it they reduce profit.

When the firm acts as a monopsonist it is constrained by the supply curve of the competitive input suppliers. It therefore maximizes profit by choosing the (p_1, z_1) combination on S_1 which yields the highest profit. This is at (p_1^*, z_1^*) where the indifference curve I^* is tangent to S_1 . Since the supply curve is positively sloped the point of tangency between I^* and S_1 must also occur where I^* is positively sloped and

$$R'_f(z_1^*) > p_1^1$$

As we saw earlier the monopsonist will maximize profit by demanding less of the input than would a firm which treated the input price as a parameter.

The effect of monopsony and output monopoly on the input market

When the output is produced from two or more inputs the analysis of the effect of both monopsony and output monopoly on the price of one of the inputs is complicated, because the use of the other input is likely to change as well, thus shifting the MRP_1 curve. If the output is produced by a single input this complication does not arise, and it is possible to show the implications of monopsony and output monopoly in a single simple diagram such as Fig. 14.5. Since there is a single input z_1 its marginal product depends only on z_1 and so the marginal revenue product MRP_1 and the value of the marginal product VMP_1 curves in Fig. 14.5 are fixed. S_1 and MBC_1 are supply and marginal buyer cost curves. There are four possible equilibria in this input market, where suppliers treat the price of

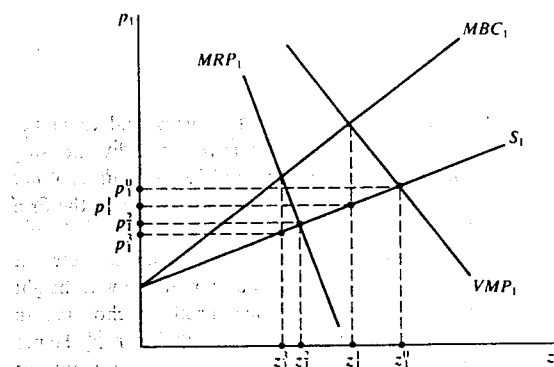


Fig. 14.5

z_1 as a parameter. If the firm also treats p_1 as given, i.e. if it acts as if it has no monopsony power and if it also treats output price as a parameter then VMP_1 is its demand curve for z_1 and the market price is p_1^0 . If the firm uses its monopsony power but continues to treat output price as a parameter it will equate VMP_1 to MBC_1 and set the price p_1^1 . If the firm monopolizes its output market but regards p_1 as a parameter its demand curve for z_1 is MRP_1 and the price of z_1 is p_1^2 . Finally, if the firm exercises both monopoly and monopsony power it equates MRP_1 and MBC_1 and sets a price p_1^3 . We see therefore that the price in an input market is reduced below the competitive level p_1^0 by both monopsony and monopoly power. The less elastic are the demand for output and the supply of input functions, the lower will be the price paid to suppliers of the input.

Exercise 14B

1. Under monopoly there is no supply curve for the monopolized output in the sense of a one to one correspondence between market price and quantity produced. Show that under monopsony there is no market demand curve for the monopsonized input.
2. Analyse the monopsonist's cost minimization problem and the monopsonist's cost curves. (Hint: what does [B.1] imply about the isocost curves?) Show that at the monopsony equilibrium the input price ratio is not in general equal to the ratio of marginal products.
3. What is the effect of minimum wage legislation on the level of employment in (a) a competitive labour market, (b) a monopsonized labour market?
4. *Discriminating monopsony.* Suppose that a monopsony employer can segment its workers into two groups (men and women) and pay the two groups different wages. Show that it will pay a lower wage to the group with the less elastic supply function. What would be the effect on employment and wages of legislation which made it illegal to discriminate in this way?
5. Show that in the multi-input case, where the firm uses n inputs to produce its output and has monopsony power only in the market for input 1, that it is possible to use indifference curves in (p_1, z_1) space to analyze its behaviour in the markets for input 1. (Hint: maximize profit for given (p_1, z_1) and then use the envelope theorem.)

C. Unions as monopoly input suppliers

We define a union as any association of the suppliers of a particular type of labour which is formed with the aim of raising wages or improving working conditions. A union need not, of course, be described as such by its members: many professional associations (such as the British Medical Association and the Law Society) act as unions. Not all unions may be successful in raising the wages of their members above the competitive level. The union, like any would-be monopolist, must be able to control the supply of labour offered

to firms. One method of doing this is to ensure that only union members can sell their labour in that particular market, a device known as the 'closed shop'. The closed shop may, by itself, reduce the supply of labour to the market if some potential workers dislike being union members as such. In general, however, the closed shop must be coupled with restrictions on the number of union members if all members are to be employed, since higher wages will increase the number of workers wishing to join the union, i.e. become employed at the higher wage.

If the union can act as a monopolist its behaviour will depend on the objective it pursues. By analogy with Chapter 13 it may be useful to distinguish between the objectives of the officials who run the union and those of the members. In the case of the firm, where conflicts of interest may exist between shareholders and managers, the extent to which the managers pursue the interests of the shareholders depends on the incentive system which relates managerial pay to profits and on the threat of product or capital market competition. Similar mechanisms may be at work in the case of the union. Officials' salaries can be related to the pay of members of the union. Unions which do not attend sufficiently closely to their members' interests may start to lose members to rival unions. Officials may be controlled directly through elections, but here the control mechanism may be much weaker than in a firm. Each union member has only one vote and so many members must cooperate to change the officials. Shareholders vote in proportion to the numbers of shares held and so a relatively small group of individual shareholders may exercise effective control.

It is by no means obvious that the political structure of a union will generate any well-defined preference ordering, let alone one which reflects the interests of its members. (See the discussion of the Arrow Impossibility Theorem in Chapter 17, section F.) However, we will assume that such a preference ordering exists and can be represented by a utility function $U(w, z)$ where w is the wage paid to union members and z is the number of union

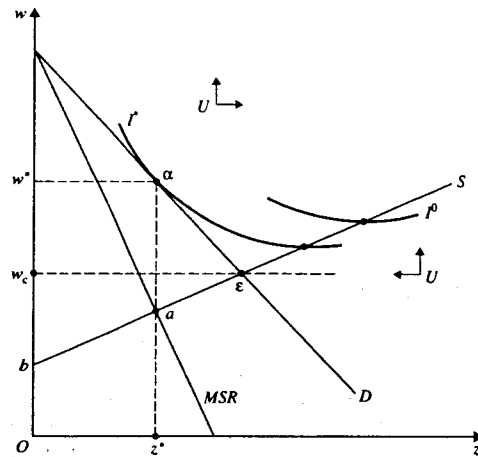


Fig. 14.6

members employed. (We assume that hours of work are fixed.) We illustrate the implications of different assumptions about union preferences by specifying three different forms for U .

The demand side of the labour market monopolized by the union is assumed to be competitive and the union is constrained to choose a wage and employment combination on the labour market demand curve D in Fig. 14.6. MSR is the corresponding 'marginal revenue to the seller' curve which shows the rate at which the total wage bill wz varies with z . S is the supply curve showing the minimum wage necessary to attract different numbers of workers into the industry. S plots the *reservation wage* or 'supply price' of workers. The competitive equilibrium in the absence of an effective union monopoly would be at ϵ with a wage rate of w_c .

The *economic rent* earned by a worker is the difference between the wage paid and the wage necessary to induce that worker to take a job in the industry. The total economic rent earned at any given wage is the difference between the total wages paid wz and the area under the labour supply curve up to the employment level. One possible objective for the union would be to maximize the total economic rent of the workers in the industry. In this case the union's utility function would be

$$U(w, z) = wz - \int_0^z \omega(\bar{z}) d\bar{z} \quad [C.1]$$

where $\omega(z)$ is the *inverse supply function* of union members, showing the wage $\omega(z)$ necessary to induce a supply of z workers. If S is interpreted as a marginal cost curve we can see that the union's problem is identical to that of a profit maximizing monopolist. Hence the union would restrict employment to where MSR cuts S by setting a wage w^* , yielding a total rent of $zabw^*$ for its members.

Equivalently, we could consider the union's indifference curves in (w, z) space. They have slope

$$\left. \frac{dw}{dz} \right|_{dU=0} = - \frac{U_z}{U_w} = - \frac{w - \omega(z)}{z} \quad [C.2]$$

Since $U_w = z > 0$ higher indifference curves are preferred to lower ones: the union is better off on I^0 than on I^* . To the left of the supply curve $w > \omega$ and so $U_z > 0$; to the right of S , where $w < \omega$, $U_z < 0$. Hence the indifference curves are \cup -shaped about the supply curve. The union is constrained by the market demand curve that it faces and chooses the wage employment combination (w^*, z^*) where its indifference curve I^* is tangent to D . (Where would the solution have been if the union takes the wage rate as a parameter which is unaffected by the employment level?)

An alternative union objective function is

$$U(w, z) = wz \quad [C.3]$$

For example, the union officials may wish to maximize the total wage bill because union membership fees are proportional to the wage and the salaries of officials or other benefits (pleasant working conditions, union conferences in exotic locations) may increase with the union's income. Since $U_w = z > 0$, $U_z = w > 0$, the indifference curves corresponding to [C.3] are negatively sloped and are rectangular hyperbolas. Higher indifference curves

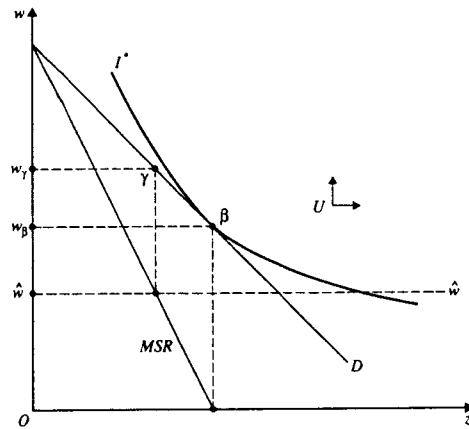


Fig. 14.7

are preferred to lower. In Fig. 14.7 the union maximizes wz by choosing the point β on D where the indifference curve I^* is tangent to D . At this point MSR is zero and the total seller revenue (the wage bill wz) is maximized.

We have so far ignored the possibility that the union may have unemployed members: at the wage set by the union not all of its members can find employment in the industry. Consider a union which has a total of z^0 members of whom z are employed at the wage w and who get a utility of $u(w)$ ($u' > 0, u'' < 0$). The $z^0 - z$ members who are not employed get unemployment pay of \hat{w} yielding utility of $\hat{u}(\hat{w})$. u and \hat{u} may be different utility functions to reflect the fact that members care about being employed or not as well as about the income they receive. Suppose that employed workers are chosen at random each period from the pool of union members, so that each member has a probability of z/z^0 of being employed and $(z^0 - z)/z^0$ of being unemployed. Assume that each member evaluates this risky prospect by its expected utility:

$$[u(w)z + \hat{u}(\hat{w})(z^0 - z)]/z^0 \quad [C.4]$$

(See Chapter 19 for a discussion of expected utility as a representation of preferences under uncertainty.) A union run in the interests of its members would aim to maximize [C.4], or since z^0 is constant, to maximize

$$U(w, z) = u(w)z + \hat{u}(\hat{w})(z^0 - z) = [u(w) - \hat{u}(\hat{w})]z + \hat{u}(\hat{w})z^0 \quad [C.5]$$

If we assume that union members are only interested in income and have a constant marginal utility of income we obtain the simple union utility function

$$U(w, z) = (w - \hat{w})z + \hat{w}z^0 \quad [C.6]$$

(The union indifference curves are now rectangular hyperbolas with a horizontal axis at \hat{w} .) Since \hat{w} and z^0 are constants, [C.5] is maximized by maximizing $(w - \hat{w})z$ and the

union's optimization problem is now analogous to that of a monopolist with a constant 'marginal cost' of \hat{w} . In Fig. 14.7 the union would choose the point γ on D , determined by the intersection of its marginal revenue curve MSR and its 'marginal cost' curve at \hat{w} .

It is possible to construct many models of the above kinds, each of which may be appropriate to a particular union or industry. A model of the way in which the union's objectives are determined is necessary in order to be able to predict what objectives will be dominant in what circumstances. This will require a detailed specification of the political constitution of the union, including the frequency and type of elections, whether officials are elected or are appointed and controlled by elected representatives and so on. In addition, the theory could be extended to take account of inter-union conflict or cooperation: will unions compete for new members? In what circumstances will unions merge or collude? It would be interesting to approach these questions using the concepts of oligopoly theory developed in Chapter 12.

Exercise 14C

1. *Rent maximization.* Confirm formally that the rent maximizing union will set a wage of w^* as shown in Fig. 14.6.
2. Sketch the indifference curves for the union utility functions [C.5] and [C.6] and the corresponding (w, z) points chosen by the union. What are the implications of assuming (a) that $\hat{u}(\hat{w}) < u(w)$ when $\hat{w} = w$ and (b) that $d\hat{u}(\hat{w})/d\hat{w} < du(w)/dw$ when $\hat{w} = w$?

D. Bilateral monopoly

Bilateral monopoly is a market situation in which a single seller confronts a single buyer. For definiteness and continuity, we consider a labour market in which supply is monopolized by a union and there is a single buyer of labour. z is the sole input in the production of an output $y = f(z)$. The revenue from sale of y is $R(f(z))$ and the MRP curve in Fig. 14.8 plots the marginal benefit to the buyer of z : $R'f' = MR \cdot MP$, using the notation of section A. The average revenue product curve ARP plots $R/z = py/z = pAP$ where AP is the average product of z : y/z .

The objective function of the firm is its profit function

$$\pi = R(f(z)) - wz = \pi(w, z) \quad [D.1]$$

and its indifference curves in (w, z) space are \cap -shaped about the MRP curve. (Recall section B.) If the firm acted as a monopsonist facing competitive labour suppliers, it would announce a wage rate at which it is willing to hire workers and employment would be determined by the supply curves of the workers.

Suppose that the union has the simple objective function

$$U(w, z) = (w - \hat{w})z + \hat{w}z^0 \quad [D.2]$$

examined in section C, where z^0 is the number of union members and \hat{w} is the income or

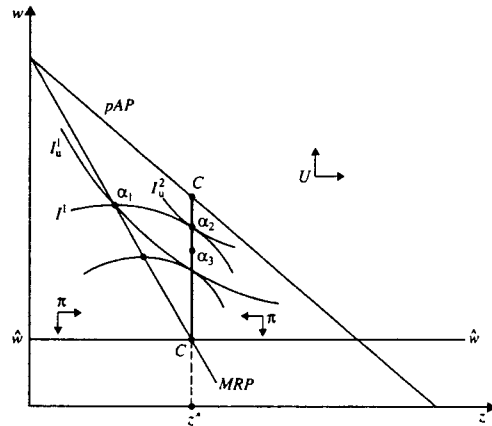


Fig. 14.8

wage of those who are unemployed. The union's indifference curves are hyperbolas, rectangular to the $\hat{w}\hat{w}$ lines, with slope

$$\left. \frac{dw}{dz} \right|_{dU=0} = -\frac{w - \hat{w}}{z} \quad [\text{D.3}]$$

If the union acted as a monopolist with respect to the labour supply of its members it would announce a wage rate at which its members would be willing to supply labour and employment would be determined by the demand curve for labour.

When a single buyer and a single seller of labour confront each other it seems implausible that either party will treat a wage rate announced by the other as parametric and passively adjust either their supply or demand. Both will realize that they possess market power in the sense that, by refusing to demand or supply labour at a wage announced by the other, they can prevent any gains from trade being achieved and thus impose costs on the other. The two parties must therefore agree on a wage and an employment level before production can occur.

We assume in this section that the agreement between the union and the firm is the outcome of a *cooperative game*. In such a game all the actions of the parties are controlled by a binding agreement between them specifying what each will do. The cooperative game approach to bargaining is concerned solely with the content of the agreement. It ignores the process of bargaining and negotiation by which agreements are reached. (We examine the alternative non-cooperative game approach, which does pay more attention to the bargaining process, in sections F and G.) We attempt to predict the agreement by requiring that it satisfy certain 'reasonable' conditions.

Two obvious conditions to impose are

1. *individual rationality*: any agreement should leave both parties at least as well off as they would be if there was no agreement;

2. *efficiency*: there should be no other agreement which would make one of them better off and the other no worse off.

If an agreement satisfies these requirements it is an *efficient bargaining solution* to the cooperative bargaining game.

Applying these conditions provides a partial answer to the question of what agreement will be reached by the union and the firm. If there is no agreement and therefore no employment, the firm will have zero profit. Any agreement which yields a (w, z) combination on or below its average revenue product curve pAP will satisfy the individual rationality constraint for the firm. If the union achieves zero utility if there is no agreement, it will be no worse off with an agreement at any point on or above the line $\hat{w}\hat{w}$. Thus the set of individually rational agreements which make both parties no worse off is the triangle bounded by the vertical axis, $\hat{w}\hat{w}$ and pAP in Fig. 14.8.

Imposing the efficiency requirement further reduces the set of possible bargains. If the parties' indifference curves intersect at a point such as α_1 it is always possible to find another point or bargain which makes at least one of them better off and the other no worse off. Thus moving from the agreement α_1 where the indifference curves I^1 and I_u^1 intersect to the agreement α_2 will make the union better off since α_2 is on the higher indifference curve I_u^2 . The firm is no worse off at α_2 since both point α_1 and α_2 are on I^1 . A move from α_1 to α_3 would make both union and firm better off.

A necessary condition for efficiency is that the parties' indifference curves are tangent:

$$\left. \frac{dw}{dz} \right|_{d\pi=0} = -\frac{\pi_z}{\pi_w} = \frac{R'f' - w}{z} = \left. \frac{dw}{dz} \right|_{dU=0} = -\frac{w - \hat{w}}{z} \quad [\text{D.4}]$$

which implies

$$R'f' = \hat{w} \quad [\text{D.5}]$$

All agreements satisfying [D.5] are efficient. Notice that [D.5] depends only on the level of employment z (which enters into $R'f'$) and not on w . The locus of points where [D.5] is satisfied and the agreement is efficient is a vertical line at z^* where MRP cuts $\hat{w}\hat{w}$.

The set of agreements satisfying individual rationality and efficiency is the *contract curve*. In the current model the contract curve has a particularly simple form: it is the line CC in Fig. 14.8 between the pAP and $\hat{w}\hat{w}$ curves where indifference curves are tangent and the parties no worse off than if they do not agree.

The efficient bargain model predicts the level of employment z^* the parties will agree on but it is unable to predict the wage rate at which the workers will be employed. This is perhaps unsurprising: the parties can agree to choose an employment level which will maximize their potential gains from agreement: the difference between the firm's revenue $R(f(z))$ and the 'cost' of labour $\hat{w}z$ as perceived by the union. A change in z which increases $R - \hat{w}z$ can make both parties better off and they can therefore agree to it. However, for fixed z , changes in the wage rate have precisely opposite effects on their utilities:

$$\pi_w = -z, \quad U_w = z$$

With z held constant changes in w merely make one party better off at the expense of the other. In Fig. 14.8 the firm will always prefer a bargain lower down CC and the union a bargain higher up CC .

One way to remove the indeterminacy of the bilateral monopoly model is to impose additional requirements on the agreement or solution of the cooperative bargaining game. We do this in the next section, using a more general notation to emphasize the wider applicability of the results.

Exercise 14D

- Derive the contract curve if the union's objective function is given by [C.2] or [C.3].
- Non-linear union utility function.** Suppose that the union has the utility function $U(w, z) = [u(w) - \hat{u}(\hat{w})]z + \hat{u}(\hat{w})\hat{z}$, where $u' > 0$, $u'' < 0$.
 - Show that contract curve has a positive slope.
 - Now suppose that the firm's production function is $y = f(z - \ell) = f(n)$ where z is total employment by the firm, ℓ is the number of workers who contribute nothing to production (they spend all day playing cards) and n is the effective labour force employed by the firm. The firm's marginal revenue function is strictly concave in output: $R'' < 0$. For outputs in excess of $f(n^0)$ its marginal revenue is negative. When employed, workers are assumed not to mind whether they play cards or produce output and they get the same wage w . The agreement between the union and the firm now specifies w , z and ℓ .
 - Show that the contract curve now has a horizontal segment for $z \geq n^0$. Interpret a bargain struck at a point on the horizontal segment.
 - What effect would an increase in the demand for the firm's output have on the contract curve? What would happen to output, employment and the wage rate if the agreement was on the horizontal segment?

E. Cooperative bargaining games*

In this section we continue the examination of the cooperative game approach to modelling bargaining. We adopt a more general framework to emphasize that bargaining games arise widely and that the same concepts can be applied in very different contexts. We consider two individuals who can make an agreement in some set of possible agreements A . In the previous section A was a set of employment and wage combinations. When two individuals are negotiating over the sale of a house A might be a set of multi-dimensional vectors, whose elements denote the price to be paid, the date of the sale, the form of payments, which of the fixtures and fittings are to be included in the sale. If the parties fail to agree the outcome is the disagreement event d . In the bilateral monopoly of section D the disagreement event is zero employment and no payment by the firm to the workers.

The bargainers have preferences over A and d which are represented by the cardinal utility functions u_i ($i = 1, 2$). In the analysis of the consumer in Chapters 3 to 5 preferences satisfied a set of axioms which ensured that they were representable by ordinal utility functions, unique up to order-preserving transformations. We are now placing stronger restrictions on preferences. If u_i and v_i both represent individual i 's preferences then v_i must be a positive linear transformation of u_i : $v_i = \alpha_i + \beta_i u_i$ with $\beta_i > 0$ (See Chapter 19 for a full discussion of cardinal utility.)

The *utility payoff set* U is the set of possible utility combinations which can be produced by an agreement a in A . Formally,

$$U = \{u_1(a), u_2(a) | a \in A\}$$

The *disagreement utilities* are $\bar{u}_i = u_i(d)$ and the *disagreement point* is $\bar{u} = (\bar{u}_1, \bar{u}_2)$. We assume that the bargainers care only about the outcome of bargaining and not about the procedure or process by which agreement is reached. Thus their attitude to agreements is fully reflected in their utilities. As far as they are concerned the bargaining situation is completely described by U and \bar{u} . We therefore attempt to predict the outcome of the cooperative bargaining game as though the parties bargained directly over utility levels, rather than over the terms of an agreement which implies particular utility levels. In many bargaining situations an agreement in terms of utility levels will imply a unique agreement in A .

A number of not very demanding requirements are placed on the utility payoff set (which implies restrictions on A , d and the individuals's preferences). It is assumed that (a) U is closed, bounded and convex; (b) $\bar{u} \in U$; (c) that there is a $u = (u_1, u_2) \in U$ such that $u_i > \bar{u}_i$ for $i = 1, 2$. The closedness requirement will be satisfied if A is closed, which it will be for most bargaining situations. Boundedness is also not demanding since it means that all elements in A yield a finite utility to both individuals. Convexity is also not restrictive. For example, it is satisfied by the bargaining situation in section D. It will always be satisfied if it is possible for the parties to choose the outcome in A by an agreed randomization rule. (See Question 1, Exercise 14E.) The assumption (b) that \bar{u} is in U merely means that the parties can agree to give themselves what they would get if there is no agreement, i.e. they can agree to disagree. Finally, (c) is necessary to ensure that the game is interesting: if there is no agreement which makes both parties strictly better off than not agreeing they would have no incentive to cooperate.

A *bargaining problem* or *cooperative bargaining game* is a utility payoff set U and a disagreement point \bar{u} satisfying the above assumptions. A *bargaining solution* is a rule which can be applied to all bargaining problems to pick a unique point in U as the outcome. Formally, a bargaining solution is a function s from the set (U, \bar{u}) to a point $(s_1, s_2) = (s_1(U, \bar{u}), s_2(U, \bar{u}))$ in U .

There are a large number of such functions. For example the rule 'maximize u_1 subject to $u \in U$ and to $u_2 = \bar{u}_2$ ' is a bargaining solution, though not an appealing one in many situations. It implies that the first party has all the bargaining strength or ability. An even more objectionable feature is apparent if we apply it to the union-firm model of section D. Labelling the union, say, as the first party is essentially arbitrary. Changing the labelling, so that the firm is the first party, would lead to a different outcome (with lower w). Thus the agreement would be dependent on arbitrary features of the model.

Nash bargaining solution

J. Nash argued that a bargaining solution ought to satisfy four reasonable requirements and then showed that there was only one rule – the Nash bargaining solution – satisfying those requirements. The Nash bargaining solution is a solution concept for *cooperative* bargaining games. It is to be distinguished from the Nash equilibrium which we have made

extensive use of in earlier chapters and which we will use again in section F. The Nash equilibrium is a solution concept used in *non-cooperative* games. In general the outcome derived by applying the two solution concepts will *not* coincide.

Nash's axioms, or reasonable requirements that the solution to a cooperative bargaining game should satisfy, are:

1. *Efficiency (E)*. There should be no feasible bargain which makes at least one party better off and the other no worse off than at the outcome chosen by the solution. Formally, there must be no $u \in U$ such that $u_i \geq s_i(U, \bar{u})$ for $i = 1, 2$ and $u_i > s_i(U, \bar{u})$ for $i = 1$ or 2 . In terms of part (a) of Fig. 14.9 the solution outcome must lie on the upper right boundary of the utility payoff sets U and V .
2. *Linear invariance (LI)*. Consider two bargaining games (U, \bar{u}) and (V, \bar{v}) where the second game is derived from the first by transforming the players' utility functions from u_i to $v_i = \alpha_i + \beta_i u_i$ ($\beta_i > 0$). The two games have exactly the same set of potential bargains A and disagreement point d and the players' preferences are the same. The only difference between the games is the numerical representation of their preferences. It seems reasonable to require that the only effect of the relabelling of the utility functions should be to relabel the solution outcome in exactly the same way: $s_i(V, \bar{v}) = \alpha_i + \beta_i s_i(U, \bar{u})$.

The invariance requirement implies that the solution outcome in A should not be affected by the numerical representation of the parties' preferences if the underlying preferences are unchanged. This is illustrated in part (a) of Fig. 14.9 where individual 1's utility is transformed from u_1 to $v_1 = \alpha_1 + u_1$. The transformation alters U to V and \bar{u} to \bar{v} . The solution outcome to the new game (V, \bar{v}) is now $s(V, \bar{v})$ but this corresponds to the same agreement in A in part (b) of the figure.

A bargaining game is *symmetric* if (a) $\bar{u}_1 = \bar{u}_2$, so that the disagreement point lies on the 45° line in (u_1, u_2) space; and (b) U is symmetric about the 45° line. Thus the game illustrated in Fig. 14.9 is not symmetric. In Fig. 14.10 only the game in part (a) is symmetric. Note that if $\bar{u}_1 \neq \bar{u}_2$ we can always transform u_1 to $u_1^* = \alpha_1 + u_1$ with $\alpha_1 = \bar{u}_2 - \bar{u}_1$ to yield a game with the disagreement point on the 45° line. From the

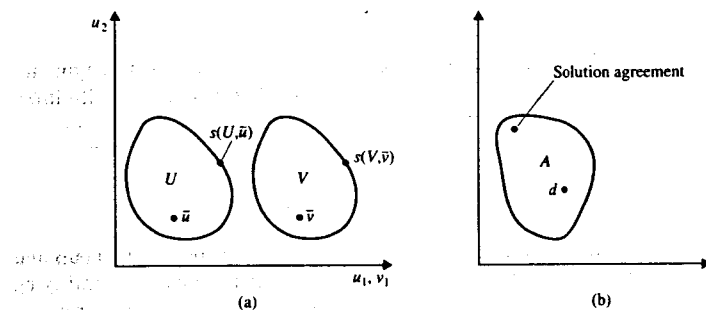


Fig. 14.9

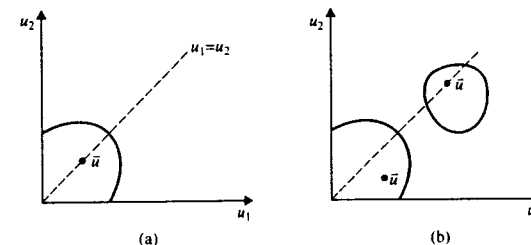


Fig. 14.10

invariance requirement, this transformation will not affect the solution outcome in any essential way. Thus the restrictive part of the symmetry definition is that U be symmetrical about the 45° line. We can now state the third of Nash's requirements:

3. *Symmetry (S)*. If the bargaining game is symmetric the solution outcome must lie on the 45° line: $s_1(U, \bar{u}) = s_2(U, \bar{u})$.

The bargaining game will be symmetric if the two players have identical preferences and are in identical circumstances. It seems reasonable to require that the bargaining solution would then give them the same utility. The symmetry requirement rules out any effect of bargaining ability on the solution. Any difference in utilities between the parties at the solution must arise from differences in their circumstances or their preferences.

4. *Independence of irrelevant alternatives (IIA)*. Consider two bargaining games: (U, \bar{u}) and (U^*, \bar{u}) where $U^* \subset U$. (The games have the same disagreement point but U^* is contained in U .) If the solution outcome $s(U, \bar{u})$ to (U, \bar{u}) is in U^* then $s(U^*, \bar{u}) = s(U, \bar{u})$: the two games must have the same solution outcome.

IIA is illustrated in Fig. 14.11 where U^* is equal to U less the shaded area. IIA requires that the bargaining outcome is unchanged if an agreement which the parties

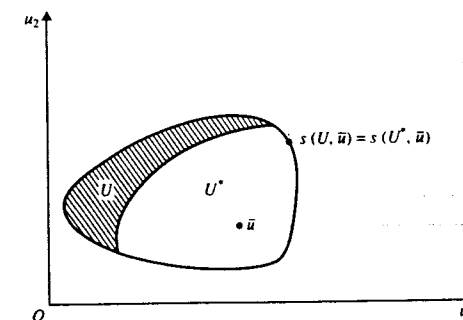


Fig. 14.11

do not make is no longer feasible. This seems reasonable if we have in mind a process of bargaining by which the parties gradually narrow down the set of potential agreements until only one remains. On the other hand, we might believe that it is plausible that the outcome ought to depend on the worst (except for disagreement) or best possible outcome for each player because these somehow reflect 'bargaining power'. The *IIA* axiom rules out solutions based on these types of intuitions.

Now consider the optimization problem

$$\max_{u_1, u_2} N(u_1, u_2) = (u_1 - \bar{u}_1)(u_2 - \bar{u}_2) \quad \text{s.t.} \quad u_i \geq \bar{u}_i \quad (i = 1, 2) \quad [E.1]$$

$$(u_1, u_2) \in U$$

derived from the bargaining game (U, \bar{u}) . The objective function in [E.1] is the *Nash product*. It is continuous and strictly quasi-concave. The assumptions about the bargaining game imply that the feasible set in [E.1] is non-empty, closed, bounded and convex. Hence applying the results in Chapter 2, section C, D and E, the solution to the optimization problem [E.1] exists and is unique. Denote the optimal values of u_1 and u_2 which solve [E.1] as $s_1^N(U, \bar{u})$ and $s_2^N(U, \bar{u})$. The *Nash bargaining solution* to the bargaining game (U, \bar{u}) is $s^N(U, \bar{u}) = (s_1^N(U, \bar{u}), s_2^N(U, \bar{u}))$.

The solution is illustrated in Fig. 14.12. The contours of the Nash product $N(u_1, u_2)$ are rectangular hyperbolas asymptotic to axes with an origin at the disagreement point \bar{u} . At s^N the contour is tangent to the upper right boundary of U . (Note the constraint $u \geq \bar{u}$ ensures that we can ignore points in the other quadrants centred at \bar{u} .) Thus the Nash bargaining solution satisfies *E*. It is also clear that it satisfies *IIA*, as the reader should check by sketching in new bargaining games (U^*, \bar{u}) with $U^* \subset U$ and $s^N \in U^*$. Since the disagreement point is unchanged the contours of N are not affected, so that s^N will maximize N over U^* . The symmetry requirement *S* is also satisfied by the Nash bargaining solution. If $\bar{u}_1 = \bar{u}_2$ the contours of N will be symmetrical about the 45° line. If U is also symmetrical about the 45° line, the solution s^N will also be on the 45° line and *S* is satisfied. (Sketch some diagrams to show this.)

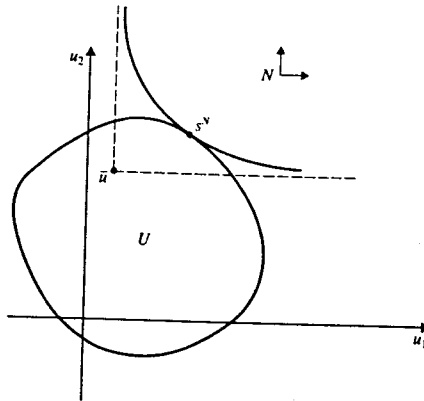


Fig. 14.12

We can show also that *LI* holds for the Nash solution. Let $v_i = \alpha_i + \beta_i u_i$ ($\beta_i > 0$) so that

$$\begin{aligned} N(v_1, v_2) &= [(\alpha_1 + \beta_1 u_1) - (\alpha_1 + \beta_1 \bar{u}_1)][(\alpha_2 + \beta_2 u_2) - (\alpha_2 + \beta_2 \bar{u}_2)] \\ &= [\beta_1 u_1 - \beta_1 \bar{u}_1][\beta_2 u_2 - \beta_2 \bar{u}_2] \\ &= \beta_1 \beta_2 N(u_1, u_2) \end{aligned} \quad [E.2]$$

The solution to the problem of maximizing the Nash product $N(v_1, v_2)$ subject to $v \in V$, $v \geq \bar{v}$ is $s^N(V, \bar{v})$. Since $s^N(U, \bar{u})$ solves [E.1]:

$$N(s_1^N(U, \bar{u}), s_2^N(U, \bar{u})) \geq N(u_1, u_2) \quad \text{all } (u_1, u_2) \in U \quad [E.3]$$

and [E.2] and [E.3] imply

$$\begin{aligned} N(\alpha_1 + \beta_1 s_1^N(U, \bar{u}), \alpha_2 + \beta_2 s_2^N(U, \bar{u})) &= \beta_1 \beta_2 N(s_1^N(U, \bar{u}), s_2^N(U, \bar{u})) \\ &\geq \beta_1 \beta_2 N(u_1, u_2) \\ &= N(\alpha_1 + \beta_1 u_1, \alpha_2 + \beta_2 u_2) = N(v_1, v_2) \end{aligned}$$

Hence $s_i^N(V, \bar{v}) = \alpha_i + \beta_i s_i^N(U, \bar{u})$ ($i = 1, 2$) as required for *LI*.

The Nash bargaining solution is simple to apply and, since it satisfies the four axioms, it has some appealing properties. More interestingly, Nash proved that the *Nash bargaining solution* $s^N(U, \bar{u})$ is the only bargaining solution satisfying the full requirements *E*, *LI*, *S* and *IIA*.

Proof. To prove the theorem we assume that some bargaining solution $s(U, \bar{u})$ satisfies the four axioms when applied to bargaining games (U, \bar{u}) and show that this implies that it is the Nash solution: $s(U, \bar{u}) = s^N(U, \bar{u})$. Suppose that for any bargaining game (U, \bar{u}) we apply linear transformations $v_i = \alpha_i + \beta_i u_i$ ($\beta_i > 0$) to both individuals' utility functions to get the transformed game (V, \bar{v}) . Since $s(\cdot)$ satisfies *LI* it must be true that

$$s_i(V, \bar{v}) = \alpha_i + \beta_i s_i(U, \bar{u}) \quad (i = 1, 2) \quad [E.4]$$

But remember that the Nash solution also satisfies *LI*, so that

$$s_i^N(V, \bar{v}) = \alpha_i + \beta_i s_i^N(U, \bar{u}) \quad (i = 1, 2) \quad [E.5]$$

Hence *LI*, [E.4] and [E.5] imply that proving that ' $s(U, \bar{u}) = s^N(U, \bar{u})$ when $s(\cdot)$ satisfies *LI*, *E*, *S* and *IIA*' is equivalent to proving that ' $s(V, \bar{v}) = s^N(V, \bar{v})$ when $s(\cdot)$ satisfies *E*, *S* and *IIA*'. The latter is easier to prove if we make the cunning choice of linear transformations:

$$\beta_i = \frac{1}{s_i^N(U, \bar{u}) - \bar{u}_i} \quad \alpha_i = -\frac{\bar{u}_i}{s_i^N(U, \bar{u}) - \bar{u}_i} = -\beta_i \bar{u}_i \quad [E.6]$$

where s_1^N, s_2^N is the Nash solution of the original game. These transformations ensure that the disagreement point of the new game is the origin:

$$\bar{v}_i = \alpha_i + \beta_i \bar{u}_i = 0 \quad (i = 1, 2)$$

Since the Nash solution satisfies *LI*, the Nash solution of the transformed game $(V, \bar{v}) = (V, 0)$ is

$$(s_1^N(V, 0), s_2^N(V, 0)) = (\alpha_1 + \beta_1 s_1^N(U, \bar{u}), \alpha_2 + \beta_2 s_2^N(U, \bar{u})) = (1, 1) \quad [E.7]$$

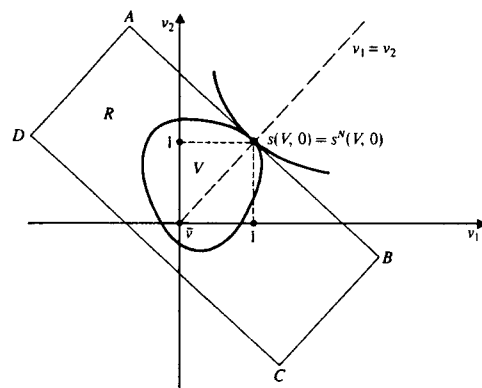


Fig. 14.13

See Fig. 14.13, where V is the utility payoff set of the transformed game, with disagreement point at the origin and Nash bargaining solution $s^N(V, 0)$ on the 45° line at $(1, 1)$. (The original game is not shown.)

We can now establish the theorem by showing that if $s(\cdot)$ satisfies E , S and IIA , applying it to the transformed game yields the Nash solution: $s(V, 0) = s^N(V, 0) = (1, 1)$. The Nash product for the transformed game is

$$N(v_1, v_2) = (v_1 - \bar{v}_1)(v_2 - \bar{v}_2) = v_1 v_2 \quad [E.8]$$

The line AB in Fig. 14.13 is tangent to the contour of the Nash product $v_1 v_2$ at the Nash solution $s^N(V, 0)$ to the transformed game. All points in V except $s^N(V, 0)$ must lie below AB , otherwise $v_1 v_2$ would not be maximized over the convex set V at $s^N(V, 0)$. (V must be convex since it is a linear transformation of the convex set U .) Since the slope of the contour of the Nash product $v_1 v_2$ at the Nash solution is $dv_2/dv_1 = -v_2/v_1 = -1$, the line AB also has slope -1 . We construct another bargaining game $(R, 0)$ with disagreement point at the origin. The utility payoff set for the new game R is the rectangle $ABCD$ which is symmetrical about the 45° line. Since U is bounded, V must also be bounded and so we can always construct a rectangle like $ABCD$ which contains the transformed game V .

Now we apply $s(\cdot)$ to the bargaining game $(R, 0)$ to get the solution outcome $s(R, 0)$. Since $s(\cdot)$ satisfies E the solution outcome must be on AB which is the upper right boundary of R . $s(\cdot)$ also satisfies S , so that the outcome must also be on the 45° line. Hence E and S imply that

$$s(R, 0) = (1, 1) = s^N(V, 0) \quad [E.9]$$

But V is a subset of R , has the same disagreement point at the origin and contains the solution outcome $s(R, 0)$. Therefore, since $s(\cdot)$ satisfies IIA , we have

$$s(V, 0) = s(R, 0) = s^N(V, 0) \quad [E.10]$$

which completes the proof.

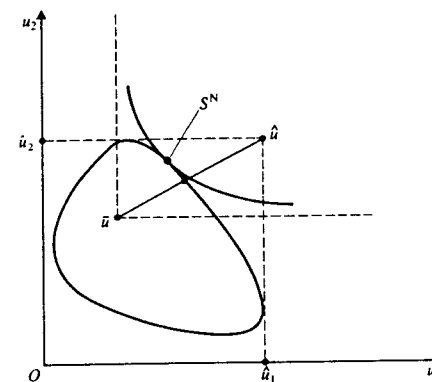


Fig. 14.14

It is remarkable that imposing four reasonable requirements on the agreement results in a unique and analytically tractable solution to the cooperative bargaining game. However, there are several difficulties with the axiomatic approach.

First, there are other, equally appealing, solutions, which satisfy different sets of axioms. For example, consider the *Kalai-Smorodinsky solution*. Denote the maximum utility that player i can get in U by \hat{u}_i . The Kalai-Smorodinsky solution is the intersection of the upper right boundary of U with a straight line connecting the disagreement point \bar{u} to $\hat{u} = (\hat{u}_1, \hat{u}_2)$. The solution satisfies E , S and LI but can differ from the Nash solution. (See Fig. 14.14.) We could think of \hat{u} as possibly reflecting the parties' bargaining strength, in that an increase in \hat{u}_i might plausibly be expected to make i less willing to accept a given agreement. As the reader should check by sketching some examples, an increase in \hat{u}_i will increase the payoff received by i in the Kalai-Smorodinsky solution, but may have no effect on the Nash solution.

Second, it is usually possible to construct examples of cooperative bargaining games in which the outcome produced by a particular bargaining solution is not intuitively appealing. (Question 2, Exercise 14E gives an example for the Nash solution.)

Third, we would usually think of an agreement between parties as resulting from a bargaining process in which offers are exchanged, rejected or modified until agreement is reached or the parties give up trying to reach agreement. The cooperative game approach leaves the process of bargaining unexamined. In particular it has nothing to say about why bargaining may not always lead to agreement. In order to model disagreement, we need to widen the set of actions being examined to include the bargaining strategies (what offers to make in what circumstances, what offers to accept or reject) that the parties adopt, as well as the actions they eventually agree upon. Since they cannot make binding agreements about their choice of bargaining strategies, we should use the tools and concepts of non-cooperative game theory for a more complete account of bargaining. The next two sections examine the approach.

Exercise 14E

1. Show that, no matter what form of A , the utility payoff set U will be convex and contain \bar{u} if the parties can choose a point in A by an agreed randomization rule which determines the probability with which each point in A is chosen.
2. *Individual rationality.* Show that if a bargain satisfies E it must also satisfy the individual rationality constraints of section D. (*Hint:* examine the definition of a bargaining game.)
3. *Divide a dollar.* Two individuals have initial wealth y_i . A mischievous dead relative has left a total of \$1 to be shared between them provided they agree on how the \$1 should be divided. If the sum of their agreed shares in the dollar is less than 1 the balance will be given to a charity. If the sum of their agreed shares is greater than 1 or if they fail to agree, they will each receive nothing and the dollar will be given to the charity. Both individuals are entirely selfish and derive no utility from the money received by the charity. Suppose that the individuals have identical utility functions $u_i = \log_e x_i$ where x_i is i 's wealth including her payout under the terms of the will.
 - (a) Draw diagrams to show the set of outcomes A and the disagreement outcome in income space and the utility payoff set. Show the effect of linear transformations of the utility functions.
 - (b) What is the Nash solution?
 - (c) Is more of the \$1 given to the richer or the poorer of the individuals?
4. *Union-firm bilateral monopoly.* Show that the union-firm model of section D satisfies the definition of a bargaining game and derive the Nash bargaining solution.
5. *Asymmetric Nash bargaining solution.* If the solution must satisfy E , LI and IIA , the bargaining solution maximizes the *asymmetric Nash product* $[u_1 - \bar{u}_1]^\alpha [u_2 - \bar{u}_2]^{(1-\alpha)}$ for some value of the parameter $0 < \alpha < 1$. α can be interpreted as relative bargaining power. Derive the asymmetric Nash solution to the union-firm bargaining problem. Show that (a) if $\alpha = \frac{1}{2}$ the asymmetric solution is identical with the Nash solution; (b) for all $0 < \alpha < 1$ the asymmetric solution has the same employment level as the symmetric Nash solution but the wage may be greater or smaller than w^* depending on α .

F. Bargaining as a non-cooperative game*

In this section we attempt to predict the outcome of bargaining between two individuals by drawing on the techniques and concepts of non-cooperative game theory. As we suggested at the end of section E, it may be fruitful to model the individuals as choosing negotiation or bargaining strategies which specify their offers and counter-offers. We can then look for an equilibrium pair of strategies and examine the agreement generated by the equilibrium strategies. Since individuals will not make binding agreements about their negotiation strategies, we seek a non-cooperative equilibrium in bargaining strategies. By contrast, in section E we derived the Nash bargaining solution for cooperative games by concentrating on the properties of agreements and made no attempt to analyse the bargaining strategies which may lead to them.

The equilibrium of a non-cooperative bargaining game depends on the exogenously given rules of the game which specify the types, sequences and timing of offers and responses to offers the parties are allowed to make, the information available to them when they make offers and responses and whether bargaining takes place over a finite or infinite time horizon. In this section we consider the Rubinstein (1982) bargaining model and show how a unique equilibrium can be derived by using the concept of subgame perfection introduced in Chapter 13, section 3. We will also indicate how the solution depends on particular features of the specification of the bargaining game and the implications of alternative assumptions.

In the Rubinstein model individuals bargain over the division of a pie (or sum of money or some other divisible object which yields utility). The units of measurement are chosen so that the size of the pie is 1. At date 0 player A makes a proposal or offer that she should get x_0 in $[0, 1]$ and individual B should get $1 - x_0$. B may accept the offer, in which case the game is over, or reject it. If B rejects the offer then, in period 1, he makes a proposal that he should get y_1 and A get $1 - y_1$. If A accepts this offer the game terminates in period 1. If A rejects B 's period 1 offer, she makes a further proposal x_2 to B in period 2, which B may again accept, terminating the game, or reject. The game will continue with players making alternating offers until one of them accepts an offer.

When the game terminates in period t , with acceptance of an offer by one of the parties which gives s to A and $1 - s$ to B , their utilities are

$$u_a = \delta_t^a s, \quad u_b = \delta_t^b (1 - s) \quad [F.1]$$

where $0 < \delta_i < 1$ is individual i 's discount factor. Both individuals are impatient: pie received at a later date is less valuable than the same amount of pie received earlier. This cost from delaying settlement provides an incentive to settle and determines the equilibrium of the model.

Both players know the discount factor of the other (and know that the other knows that they know). At each date both players also know the full history of the game (the offers made and rejected) up to that date. Thus the game is one of full information.

Before the game begins individual A formulates a bargaining strategy σ_a which is a complete specification of what offers she will make in even numbered periods 0, 2, 4, ... and what offer from individual B she will accept in odd numbered periods 1, 3, 5, We can describe A 's strategy by $\sigma_a = (x_0, r_{1a}, x_2, r_{3a}, \dots)$ where x_t is A 's offer in even periods and r_{ta} her rejection rule: in an odd-numbered period t she turns down any proposal by B which gives her less than r_{ta} . B 's strategy σ_b is analogously defined: $\sigma_b = (r_{0b}, y_1, r_{2b}, y_3, \dots)$. The strategies specify offer or accept/reject decisions in period t which depend on the whole history of the game up to period t , i.e. on the whole sequence of offers and counter offers in periods 0 to $t - 1$. (Thus x_t, r_{ta}, y_t, r_{tb} should be written as functions of the history of the game up to $t - 1$.) Despite the fact that the strategies could be extremely complex, Rubinstein proved that the players' equilibrium strategies are in fact simple and result in the game finishing in the first period with player A making an offer which is accepted by B .

Constant strategy equilibria

The game structure is essentially stationary in that, if the game has not terminated by period t , it has the same structure as the game at period $t - 2$: the size of pie is unchanged,

the possible offers and accept/reject rules are the same at t and $t - 2$ and the parties' preferences concerning pie at the current date and in the future are the same. A has a *constant strategy* if she always makes the same proposal \bar{x} when it is her turn to make offers and if she always rejects proposals from B unless they give her at least \bar{r}_a . We write constant strategies as $\bar{\sigma}_a = (\bar{x}, \bar{r}_a)$. B 's constant strategies $\bar{\sigma}_b = (\bar{y}, \bar{r}_b)$ are similarly defined. The stationarity of the game suggests that we can restrict ourselves to looking for constant strategy equilibria. (In the last sub-section we prove that this restriction is justified because the only sensible equilibrium involves constant strategies.)

One obvious equilibrium concept for this non-cooperative game is the *Nash equilibrium*. Recalling Chapter 12, section B, a Nash equilibrium (NE) is a pair of strategies (σ_a^*, σ_b^*) such that σ_a^* is the best reply for A to σ_b^* and vice versa. (Remember that the Nash equilibrium is a concept which is defined only for non-cooperative games and should not be confused with the Nash bargaining solution of section E which is defined only for cooperative bargaining games.) Unfortunately, there are many NE in the current non-cooperative bargaining game, even though we restrict the players to constant strategies.

For example, consider *intransigent* constant strategies defined by $\bar{\sigma}_a = (\bar{x}, \bar{x})$ and $\bar{\sigma}_b = (\bar{y}, \bar{y})$. Here player A always proposes \bar{x} in periods 0, 2, 4, ... and accepts any proposal y where $(1 - y) \geq \bar{x}$ in periods 1, 3, 5, ... In effect A insists in getting at least \bar{x} of the pie whenever the game terminates. Similarly, B always proposes \bar{y} in periods 1, 3, 5, ... and accepts any proposal x where $(1 - x) \geq \bar{y}$ in periods 0, 2, 4, ...

A pair of intransigent strategies need not constitute a NE: if $\bar{x} + \bar{y} > 1$ then one party always rejects the other's proposal and if $\bar{x} + \bar{y} < 1$ then either party would be better off increasing their intransigent demand. However, any pair of *reciprocal* intransigent strategies with $\sigma_a = (\bar{x}, \bar{x})$, $\sigma_b = (1 - \bar{x}, 1 - \bar{x})$ is a NE. This pair of strategies results in the game terminating in period 0 when B accepts A 's proposal, yielding A utility of \bar{x} and B a utility of $(1 - \bar{x}) = \bar{r}_b = \bar{y}$.

To see that reciprocal intransigent strategies are a NE consider whether A can do better by choosing another strategy, given that B will continue to choose his intransigent strategy. For example, consider the reciprocal intransigent strategies $\bar{x} = 0.6$, $\bar{r}_a = 0.6$ and $\bar{y} = 0.4$, $\bar{r}_b = 0.4$. Suppose that, at date 0, A makes an offer which yields B only 0.3. Since he sticks to his intransigent strategy, B rejects this proposal because $0.3 < \bar{r}_b = 0.4$. Given that B sticks to his intransigent strategy in all periods, the game can only terminate when A makes an offer acceptable to B (which requires $(1 - x) = y \geq 0.4$, i.e. $x \leq 0.6$) or when A accepts the offer from B (giving her a slice of pie $(1 - 0.4) = 0.6$). Thus by demanding more than 0.6 in period 0, A , at best, merely postpones receipt of the same amount of pie by at least one period, yielding utility of at most $0.6\delta_a < 0.6$. Hence, given that B sticks to his intransigent strategy, A cannot do better by deviating from her intransigent strategy in period 0.

Similar arguments apply to any other deviation by A or by B from any reciprocal intransigent strategy. Since neither can do better by unilaterally deviating from any reciprocal intransigent strategies, the NE is clearly not unique: all reciprocal intransigent strategies are NE.

Reciprocal intransigent strategies are not intuitively appealing NE. They require that A believes that B sticks to his intransigent strategy in the face of a deviation by A (and vice versa). But B 's intransigence is not credible: faced with a deviation by A he could do better if he does not stick to his intransigent strategy. Thus suppose that the reciprocal

intransigent strategies are $\bar{x} = 0.6$, $\bar{r}_a = 0.6$, $\bar{y} = 0.4$, $\bar{r}_b = 0.4$ and that B 's discount factor is $\delta_b = 0.7$. B 's intransigent strategy requires him to refuse all offers from A which give him less than 0.4 and to make an offer of 0.4 whenever it is his turn to do so. Suppose that, at date 0, A proposes $x = 0.7$, which would give B a slice of pie of 0.3 in period 0. In all other periods 1, 2, ... A 's strategy is the same as her intransigent strategy. If B sticks with his intransigent strategy he will reject the proposal. The game will then terminate when A accepts B 's proposal of $y = \bar{y} = 0.4$ in period 1. By sticking to his intransigent strategy in period 0, B will get a slice of pie of size 0.4 in period 1 which gives him a discounted utility of

$$\delta_b \bar{y} = (0.7)(0.4) = 0.28 < 0.3$$

Thus he does worse by sticking to his intransigent strategy than by accepting A 's proposal which would give him 0.3 of the pie in period 0. B 's threat to reject A 's offer is therefore not credible.

More generally: B 's threat to stick to his intransigent strategy (\bar{y}, \bar{y}) is not credible against a proposal by A of $\bar{x} + \varepsilon$, where $0 < \varepsilon < (1 - \delta_b)\bar{y}$. B would always prefer to deviate from his intransigent strategy to accept the proposal since $\bar{y} - \varepsilon > \delta_b \bar{y}$. Similar reasoning establishes that A 's threat to stick to her intransigent strategy, whatever proposals are made by B , is also not credible.

Recall from Chapter 12, section D that strategies constitute a *subgame perfect equilibrium* (SPE) if they constitute NE at every stage or subgame reached in the game. SPE strategies therefore cannot involve incredible threats because such threats require behaviour which is not a best reply to the other player's actions. We have just shown that, since all reciprocal intransigent strategies involve incredible threats, no reciprocal intransigent strategy pair can be a subgame perfect equilibrium (SPE). We will now demonstrate that there is in fact a unique constant strategy SPE for the Rubinstein bargaining game and that in it neither player is intransigent: $\bar{x} \neq \bar{r}_a$, $\bar{y} \neq \bar{r}_b$.

We seek $\sigma_a = (\bar{x}, \bar{r}_a)$, $\sigma_b = (\bar{y}, \bar{r}_b)$ which are best replies to each other and such that player i 's refusal to accept less than \bar{r}_i is credible. If B turns down an offer of $(1 - x)$, then given A 's constant strategy, the best he can do is to make a proposal next period which gives A the least she will accept: \bar{r}_a . Since A will accept this proposal, B will get $(1 - \bar{r}_a)$ next period. It is therefore credible for B to reject $(1 - x)$ if $(1 - x) \leq \delta_b(1 - \bar{r}_a)$. Thus B 's credible rejection criterion must satisfy $\bar{r}_b \leq \delta_b(1 - \bar{r}_a)$. It cannot be optimal for him to set $\bar{r}_b < \delta_b(1 - \bar{r}_a)$ since he would be worse off rejecting, rather than accepting, a proposal x such that $\bar{r}_b < (1 - x) < \delta_b(1 - \bar{r}_a)$. Hence, given A 's strategy (\bar{x}, \bar{r}_a) , B 's optimal credible rejection criterion is

$$\bar{r}_b = \delta_b(1 - \bar{r}_a) \quad [F.2]$$

Similar reasoning for A gives her credible and optimal rejection criterion

$$\bar{r}_a = \delta_a(1 - \bar{r}_b) \quad [F.3]$$

Since player i will accept \bar{r}_i it is not optimal for the other player to make a larger offer and so the offer must satisfy

$$(1 - \bar{x}) = \bar{r}_b \quad [F.4]$$

$$(1 - \bar{y}) = \bar{r}_a \quad [F.5]$$

An SPE must satisfy [F.2] to [F.5] and, since there is only one set of $(\bar{x}, \bar{r}_a, \bar{y}, \bar{r}_b)$ satisfying

these four equations, [F.2] to [F.5] define the unique SPE in constant strategies. Solving [F.2] and [F.3] for \bar{r}_a and \bar{r}_b and substituting in [F.4] and [F.5], the unique SPE in constant strategies is

$$\begin{aligned} x^* &= \frac{1 - \delta_b}{1 - \delta_a \delta_b}, & r_a^* &= \frac{\delta_a(1 - \delta_b)}{1 - \delta_a \delta_b} = \delta_a x^* \\ y^* &= \frac{1 - \delta_a}{1 - \delta_a \delta_b}, & r_b^* &= \frac{\delta_b(1 - \delta_a)}{1 - \delta_a \delta_b} = \delta_b y^* \end{aligned} \quad [\text{F.6}]$$

The game will terminate in period 0 when Player A proposes x^* , which player B accepts since $(1 - x^*) = r_b^*$. The utilities of the players at this SPE are

$$u_a^* = x^* = (1 - \delta_b)/(1 - \delta_a \delta_b), \quad u_b^* = r_b^* = \delta_b(1 - \delta_a)/(1 - \delta_a \delta_b) \quad [\text{F.7}]$$

Notice that B 's utility is not equal to his proposal but to his rejection criterion, since he accepts A 's just-acceptable offer and never gets to make a proposal.

Specification of the model

The solution to the Rubinstein bargaining model depends on the parties' discount rates because the fact that they value future consumption less than current consumption means that delay in settlement is costly. It is this which gives them an incentive to agree rather than to continue bargaining. The assumption on preferences, that consumption is discounted at a constant proportional rate, is quite strong. However, though it is necessary to get the particularly elegant form of the solution in [F.6], much weaker restrictions on preferences over future consumption will still yield a unique SPE solution with constant strategies.

Some of the assumptions about the exogenously given structure of the bargaining game are less innocuous. For example, the assumption that the parties make alternating offers, rather than one party making offers, is crucial. If only player A made proposals which player B could only accept or reject then the unique SPE has A getting all the pie. (See Question 2, Exercise 14F.) In the alternating offers game, the fact that B can credibly threaten to turn down some offers from A and make a counter proposal limits A 's ability to extract all the pie.

The alternating offers assumption is also important in that it gives a *first mover advantage* to A . If the players are identical in that they have the same discount rate: $\delta_a = \delta_b = \delta$, the SPE utilities are

$$u_a^* = \frac{1 - \delta}{1 - \delta^2} = \frac{1 - \delta}{(1 - \delta)(1 + \delta)} = \frac{1}{1 + \delta}, \quad u_b^* = \frac{\delta}{1 + \delta}$$

Thus even if the parties have identical preferences they do not receive equal utility in the SPE. The first mover A gets a larger share of the pie and the first mover advantage increases as the common discount factor becomes smaller, i.e. as the players discount the future more heavily. Since in many bargaining situations the order of play will be arbitrary the first mover advantage is a serious drawback.

To see the intuition underlying the first mover advantage note that if the game lasted

one period only and A moved first she could get all the pie, since B could not credibly threaten to turn down any offer. In the infinite horizon game A 's first mover advantage is constrained by B 's credible threat to turn down her proposal and to make a counter proposal. But B 's threat is limited, first by the fact that he has to defer receipt of the pie to exercise his threat and, second, by A 's ability to credibly turn down his proposal. Referring to [F.6] we see that A 's advantage from moving first is larger the smaller is B 's discount factor and the greater is A 's. Indeed as δ_a approaches 1 with δ_b held constant A 's share approaches 1.

Intuition suggests, correctly, that if the reduction in utility from deferring consumption by one period becomes smaller, so that delay becomes less important, the first mover advantage should disappear. K. Binmore has demonstrated this by supposing that the periods get shorter whilst preferences are essentially unchanged. (See Binmore and Dasgupta, 1987.) Imagine that the period (initially a day) is split into n shorter periods. Let δ_i continue to be the daily discount factor and let δ_{in} denote the discount factor for a period of $(1/n)$ th of a day. If player i is to have the same attitude to pie in one day's time it must be true that $(\delta_{in})^n = \delta_i$ or $\delta_{in} = (\delta_i)^{1/n}$. Inserting δ_{in} in [F.6] gives the SPE strategies as functions of the length of period $(1/n)$. Taking the limit as $n \rightarrow \infty$ (see Question 3, Exercise 14F), the shares of the players tend to

$$x^* = \frac{\log \delta_a}{\log \delta_a + \log \delta_b}, \quad y^* = \frac{\log \delta_b}{\log \delta_a + \log \delta_b} \quad [\text{F.8}]$$

If the parties have identical preferences (one-day discount factors) the SPE yields an intuitively appealing equal division of the pie. A now has no first mover advantage because delay for one period imposes a vanishingly small cost.

Although we have emphasized that the Nash bargaining solution is applicable only to cooperative games, it is interesting to note the relationship between [F.6] and the division of the pie produced by the asymmetric Nash bargaining solution in the corresponding cooperative bargaining game (see Question 5, Exercise 14E). The asymmetric Nash solution to the cooperative bargaining game in which A and B seek to divide a pie of size 1 is found by maximizing the asymmetric Nash product

$$x^\alpha y^{(1-\alpha)}$$

subject to $0 \leq x = (1 - y) \leq 1$, where $0 < \alpha < 1$ is usually interpreted as a measure of A 's bargaining power. The amounts of pie given to the parties at the asymmetric Nash solutions are

$$x^* = \alpha, \quad y^* = 1 - \alpha \quad [\text{F.9}]$$

Comparing [F.9] and [F.8], we see that if $\alpha = \log \delta_a / (\log \delta_a + \log \delta_b)$ the shares implied by the asymmetric Nash bargaining solution to the cooperative game and the limit of the Rubinstein solution to the non-cooperative bargaining game are identical.

Proof that the unique SPE has constant strategies

Our analysis above assumed that we need only examine constant strategies in seeking a SPE. We found a unique SPE [F.6] in constant strategies. We now demonstrate that this

is the only SPE even when the players are not restricted to constant strategies. The proof is in two stages. We show that (a) the utilities the players will get in any SPE must be the same as they would get in the unique SPE in constant strategies and (b) these constant strategies are the only means of achieving these payoffs in an SPE.

(a) Each SPE gives player some piece of pie. An SPE has the characteristic that its strategies also lead to a SPE for sub-games starting in any subsequent period. The sub-game starting in period 2 has exactly the same structure as the game starting in period 0. The pie has the same size, the same player (A) will make the offer, the players' possible choices are the same. They have the same attitude to delays in consumption of pie: at dates 0 and 2 player i is indifferent between consuming s now and $\delta_i^2 s$ in t periods time and in period 2 she places utility s on consumption of s in period 2. Thus at date 2 the highest utility that player A can get from any SPE of the sub-game starting in period 2, is equal to the highest utility she could get from any SPE of the game starting in period 0. Let us call this M . Consider player B 's proposal in period 1. Since A will accept any period 1 proposal which gives her at least $\delta_a M$, B cannot do worse in any SPE than he gets by making the proposal $1 - \delta_a M$ in period 1.

Consider player A 's proposal in period 0. Since B can always reject a proposal and then make the acceptable proposal $1 - \delta_a M$ in period 1, the best that A can do in period 0 is to make a proposal giving B the share $\delta_b(1 - \delta_a M)$ and herself $1 - \delta_b(1 - \delta_a M)$. But by definition M is the highest utility that A can get from any SPE. Hence $M = 1 - \delta_b(1 - \delta_a M)$ or

$$M = (1 - \delta_b)/(1 - \delta_a \delta_b) \quad [\text{F.10}]$$

Next we define m as the lowest utility that A can obtain from any SPE and apply similar reasoning. At date 1 B knows that A can get at least m in period 2 and so B cannot do better than make a proposal giving A the share $\delta_a m$ and himself $1 - \delta_a m$. At date 0 A realizes that B will accept any proposal which gives him at least $\delta_b(1 - \delta_a m)$ and so A can assure herself of at least $1 - \delta_b(1 - \delta_a m)$ in any SPE of a game starting at period 0. But this is the definition of m and so $m = 1 - \delta_b(1 - \delta_a m)$ or

$$m = (1 - \delta_b)/(1 - \delta_a \delta_b) \quad [\text{F.11}]$$

Since, from [F.10] and [F.11], the best that A can get from any SPE is equal to the least she can get from any SPE, we see that A gets the same utility $(1 - \delta_b)/(1 - \delta_a \delta_b)$ from all SPEs. B must therefore also get the same utility $\delta_b(1 - \delta_b)/(1 - \delta_a \delta_b)$ in all SPEs. Note that these are the utilities [F.7] at the unique SPE in constant strategies.

(b) Now we establish that the only pair of SPE strategies which give the players these unique utility payoffs is the constant strategy pair [F.6]. Suppose that there is an SPE in which the first offer by A is rejected, in which case there must be a settlement reached t periods later. But such a settlement must give the players consumption levels at date t which yield the required SPE utilities when discounted back to period 0. Thus each player's consumption of pie at date t must exceed their consumption at date 0. But this is impossible since their total consumption at date 0 was just equal to the total pie. Thus there can be no SPE in which the first offer is rejected. If the first period offer is to be accepted and to yield the unique SPE utilities, A must offer the x^* and B adopt the rejection rule r_b^* defined in [F.6] in period 0. But what about their offers and rejection rules in period 1, 2, ...? An SPE must lay down strategies which constitute an SPE at all sub-games starting in periods

1, 2, In an even-numbered period it is A 's turn to make the offer and, if the strategies in this sub-game are to be SPE, they must also have the offer in the first period of the sub-game being accepted. Hence an SPE strategy for the game starting at period 0 must say that x^* and r_b^* would be chosen in all even-numbered periods.

In odd-numbered periods it is B 's turn to make the offer. The sub-games starting in odd-numbered periods must also be SPE. We can apply the argument in (a) above to show that in an alternating offer game in which B moves first, the SPE utilities must be

$$u_a^* = \delta_a(1 - \delta_b)/(1 - \delta_a \delta_b), \quad u_b^* = (1 - \delta_a)/(1 - \delta_a \delta_b)$$

Now just adapt the argument of the previous paragraph to show that all sub-games starting in odd-numbered periods must terminate in the first period with B offering y^* and A using the rejection rule r_a^* and that this gives the required SPE utilities.

Hence we have established that the only SPE strategies which yield the parties the required SPE utilities have them choosing x^* , r_b^* in all even periods and r_a^* , y^* in all odd numbered periods. Thus the only SPE strategies are the constant strategies [F.6].

Exercise 14F

1. *Finite horizon bargaining game.* Consider an alternating offer bargaining game which has the same structure as the one described in the text except that there are a finite number T of bargaining periods. If agreement is not reached by the end of period T the game terminates with neither player getting any pie. Show that the unique SPE of this game tends to the SPE of the finite horizon bargaining game as $T \rightarrow \infty$. (Hint: work backwards from period T . What is the unique SPE of the single period game starting at date T ? Given this, what is the SPE of the two period game starting at $T - 1$? Carry on back to period 0. Now take the limit.)
2. *One sided offers.* Suppose preferences are the same as those in the text but that only A makes offers in each period and B 's actions are limited to accepting or refusing them. What is the sub-game perfect equilibrium if (a) there is only one period? (b) there is a finite number of periods? (c) the number of periods is infinite?
3. Prove the assertion in the text that as the length of periods tends to zero the utility of A tends to x^* . (Hint: use L'Hopital's rule; recall that $dz^{f(n)}/dn = z^{f(n)} f'(n) \log z$.)

G. Delay and disagreement in bargaining*

Bargainers do not always settle immediately, even though delay is costly, and they sometimes fail to reach any agreement, even though there are bargains which would make both of them better off. The bargaining models of sections E and F do not provide any insight into these two common phenomena. The models cannot explain, for example, why strikes occur, why international trade negotiations can take years and why some litigants fail to reach an out of court settlement to avoid the expense of a trial. Models which use cooperative game concepts, as in section E, seek to predict the terms of agreement by

requiring them to satisfy certain axioms. Such models *assume* that the parties will always agree and so cannot be used to explain disagreement. The Rubinstein non-cooperative game model of section F does not assume that agreement will occur, but it *predicts* that the parties will always make a bargain, and do so in the first period.

What is required is an explanation of why rational players fail to realize potential gains from trade. So far in this chapter we have assumed that the parties have complete information about the preferences, opportunities and information of the other player. In this section we present a non-cooperative model in which the parties have different information and consequently may fail to agree or do so only after a costly delay.

The model, based on Fudenberg and Tirole (1983) is a simplified version of the models considered earlier in the chapter in which a buyer and seller can realize gains from trade. A seller S and a buyer B negotiate about the sale of an asset by S to B . Both B and S know that the asset has no value to S if she does not sell it to B . The value of the asset to the buyer is $b > 0$. If b was known to both parties the situation would be similar to those considered in previous sections where the parties bargain over splitting the known gain from trade b (a pie of known size) by agreeing on a price p which gives S a gain of p and B a gain of $b - p$.

Instead we assume that B may have either a high valuation ($b = h$) or a low valuation ($b = \ell$) of the asset with

$$h > \ell > 0 \quad [G.1]$$

Whatever the buyer's valuation there are always potential gains from trade. The seller does not know which type of buyer she faces but she does know that the probability of the buyer being type h is q .

The bargaining framework differs from the infinite horizon, alternating offers game of section F. There is a finite number of periods and only the seller makes proposals concerning p . The buyer can only accept or reject the proposed price. The fact that there are a finite number of periods enables us to apply backward induction reasoning to determine the parties' optimal bargaining strategies. The assumption that only the seller makes proposals about p means that the game is a form of monopoly. As we will see, the seller's incomplete information about the buyer's valuation of the asset reduces her ability to exploit her monopoly position. (Question 6, Exercise 14G extends the analysis to allow both parties to make proposals.)

One period of bargaining

If there is only one bargaining period the solution is straightforward. B 's optimal strategy depends on his type: he accepts the offer if $b \geq p$ and rejects it if $b < p$. The seller knows that this is B 's strategy and so realizes that the probability of the offer being accepted is 1 if $p \leq \ell$ (since then both type B 's would accept), q for $\ell < p \leq h$, and 0 if $p > h$. Setting $p < \ell$ cannot be optimal for S , since such an offer is certain to be accepted and S could do better by setting $p = \ell$, which is also certain to be accepted. Setting $p > h$ also is not optimal for S , since there will be no sale and she would do better by setting $p = \ell$ and getting ℓ for sure. Finally, $\ell < p < h$ cannot be optimal either, since raising p to h does not affect the sale probability. Thus the seller either sets a price of ℓ or h . She prefers the alternative which yields the greater expected value and so chooses $p = h$ if $qh > \ell$.

There is a critical probability

$$\bar{q} \equiv \ell/h \quad [G.2]$$

such that if $q > \bar{q}$ the seller is *strong* and sets a high price and if $q < \bar{q}$ the seller is *weak* and chooses the low price.

When S chooses the high price she gets an expected payoff of qh . The high price gives both types of buyer a zero surplus: the type ℓ because he does not buy at the price h and the type h because the price paid is equal to his valuation of the asset. Thus the expected combined surplus of the seller and buyers is qh . If the seller sets a low price she would get a surplus of ℓ , the type ℓ buyer would have a zero surplus and the type h buyer would get a surplus of $h - \ell$. The combined surplus if the low price is set is $\ell + q(h - \ell) = qh + (1 - q)\ell > qh$. One period bargaining with the uninformed seller setting the price may therefore result in the total surplus of the parties not being maximized. (See Question 2, Exercise 14G.)

With a single period the analysis is very similar to monopoly: the seller faces a downward sloping 'demand' curve in that the expected quantity sold declines with the price she sets. (The analogy is even closer if there is a continuous distribution of buyer types – see Question 1, Exercise 14G.) Her market power from being able to make a single take it or leave it offer is tempered by her lack of knowledge of the type of buyer she confronts. If she set a price of h , the buyer's behaviour – his response to her offer – will provide her with information on his type since it is only optimal for a type h to accept her offer. This information is, however, of no use to the seller in a one-period model: she makes her only decision – a binding offer to sell at price p – before she observes the potentially informative decision of the buyer.

Two periods of bargaining

With more than one period the game becomes much more complicated because both parties will realize that B 's responses to offers convey information. We assume that there are two periods and that both parties apply the same discount factor δ to second period income.

At date 0 the seller makes an offer p_0 . The buyer, who may be type h or ℓ , decides to accept or reject the period 0 offer. The period 0 decision rule of a type b buyer is described by the probability $a_0(p_0, b)$ that he accepts the offer p_0 . For example if a type ℓ buyer decides that he will accept all offers with $p_0 \leq \ell$ and reject all $p_0 > \ell$, his decision rule is $a_0(p_0, \ell) = 1$ for $p_0 \leq \ell$, $a_0(p_0, \ell) = 0$ for $p_0 > \ell$.

If her period 0 offer is rejected the seller makes another offer in period 1: p_1 . B 's response to the period 1 offer is again described by the probability with which he chooses to accept it: $a_1(p_1, b)$. Neither buyer nor seller can make binding commitments to take specified actions: for example in period 0 the seller cannot commit never to reduce her period 1 offer below h . (Question 5, Exercise 14G examines the implications of commitment.)

The seller formulates her strategy to maximize her expected discounted income given her probability beliefs about the type of buyer she faces. At date 0 she attaches probability q to B being type h . At date 1 she has acquired information from B 's response to her period 0 offer. If her offer p_0 was accepted the information is of no value because the game

has terminated, but if it was rejected she can use the additional information to revise her probability beliefs about the type of buyer she faces. We assume that she uses Bayes's Rule (see the Appendix to this chapter) to update her beliefs. We denote S 's updated probability that the buyer is type h given that the period 0 offer was rejected by $q_1 = q_1(p_0)$. The notation $q_1(p_0)$ indicates that the information conveyed by a rejection may vary with the offer rejected.

The equilibrium concept we use is the *perfect Bayesian equilibrium* (PBE), which is a refinement of the sub-game perfect equilibrium used in section F. A pair of strategies is a PBE if they constitute best replies to the other player's strategy in all sub-games, given that in each sub-game each player will use the past behaviour of the other and their knowledge of the other's strategy to update their probability beliefs using Bayes's Rule.

For a PBE the strategies of the players must be equilibrium strategies for any sub-game. The seller, by her initial offer p_0 , can force the game into the sub-game starting with B 's response to that particular initial offer. She chooses which sub-game will be played and will, in equilibrium, choose the sub-game which maximizes her expected discounted payoff. We therefore characterize the PBE by first working out the equilibrium strategies for the sub-games starting with a given p_0 . The equilibrium for the whole game starting with S 's choice of p_0 is then found by looking for the p_0 which maximizes her expected discounted payoff. We will only consider the case in which the seller would be strong in a one-period game: her prior probability belief q exceeds the critical value \bar{q} defined by [G.2]. (Question 3, Exercise 14G examines the weak seller case in which $q < \bar{q}$.)

The type ℓ buyer's optimal strategy is simple: in any period always refuse any price higher than ℓ and accept in any period any offer which is less than or equal to ℓ . (Buyers realize that it cannot be optimal for S to set p_1 less than ℓ in period 1 and so the type ℓ buyer cannot gain by rejecting $p_0 \leq \ell$ in the hope of getting a lower price in period 1.) Thus the type ℓ buyer's decision rules are

$$a_\ell(p_t, \ell) = 1 \quad \text{for } p_t \leq \ell, \quad a_\ell(p_t, \ell) = 0 \quad \text{for } p_t > \ell \quad (t = 0, 1) \quad [\text{G.3}]$$

The type h buyer's strategy in period 1 is also simple: since his is the last move in the game, he just compares S 's offer p_1 with his valuation of the asset and accepts if $p_1 \leq h$:

$$a_1(p_1, h) = 1 \quad \text{for } p_1 \leq h, \quad a_1(p_1, h) = 0 \quad \text{for } p_1 > h \quad [\text{G.4}]$$

In the last period of the game the seller is in essentially the same position as in the single period game *except that her probability beliefs may have been changed by the buyer's rejection of her first period offer*. We can apply the same kind of argument as in the single period model to conclude that S will set $p_1 = \ell$ or $p_1 = h$. We can therefore describe her decision rule in period 1 by the probability

$$\lambda = \Pr[p_1 = \ell]$$

Given her updated probability belief following the buyer's rejection of her first period offer p_0 , the seller's optimal second period decision rule is

$$\begin{aligned} \lambda &= 0 & \text{if } q_1(p_0) > \bar{q} \\ \lambda &= 1 & \text{if } q_1(p_0) < \bar{q} \\ \lambda &\in [0, 1] & \text{if } q_1(p_0) = \bar{q} \end{aligned} \quad [\text{G.5}]$$

S must now find a period 0 acceptance rule for the type h buyer which together with strategies [G.3], [G.4] and [G.5] yields equilibria of the sub-games starting from a first period offer. First, consider the sub-game starting with a first period offer of $p_0 = \ell$. The type h 's optimal response is clearly to accept this offer since he cannot do better by waiting even if the seller's second period strategy [G.5] resulted in $p_1 = \ell$ for sure:

$$h - \ell > \delta(h - \ell)$$

Thus he will set $a_0(\ell, h) = 1$. This acceptance rule and the decision rules [G.3], [G.4], [G.5] generate an equilibrium in the sub-game starting at $p_0 = \ell$. If the seller sets $p_0 = \ell$, the game will end after one period since both types of buyer will accept the offer, and the seller will get an income of ℓ .

Next, consider sub-games starting with S setting $p_0 > \ell$. The type ℓ buyer, following [G.3], will reject this offer whatever he believed about the period 1 offer. If the type h buyer was sure that $p_1 = \ell$, he would accept $p_0 > \ell$ if and only if

$$h - p_0 \geq \delta(h - \ell)$$

or, equivalently, if and only if

$$p_0 \leq \bar{p}_0 \equiv (1 - \delta)h + \delta\ell \quad [\text{G.6}]$$

Over the range of sub-games starting with $p_0 \in (\ell, \bar{p}_0]$ the acceptance rule $a_0(p_0, h) = 1$ and the decision rules [G.3], [G.4] and [G.5] constitute an equilibrium. If the offer $p_0 \leq \bar{p}_0$ is rejected the seller knows that such behaviour is optimal only for the type ℓ buyer and will revise her probability belief to $q_1(p_0) = 0$. Hence, from [G.4], she will set $p_1 = \ell$ and the asset will be bought by the type ℓ in the second period. If the seller faced the type h buyer the offer $p_0 \leq \bar{p}_0$ would be accepted in the first period, since the type h buyer gains nothing by delaying acceptance.

The seller's expected proceeds from choosing an initial price in the range $(\ell, \bar{p}_0]$ is $qp_0 + \delta(1 - q)\ell$ which is increasing in p_0 . (Remember that when the buyer chooses p_0 she must use the prior probability q since she only acquires information from observing the buyer's response to her initial offer.) If the seller's optimal initial offer is in the range $(\ell, \bar{p}_0]$ it must be equal to \bar{p}_0 and will yield her an expected discounted income

$$\begin{aligned} v(\bar{p}_0) &= q\bar{p}_0 + \delta(1 - q)\ell = q[(1 - \delta)h + \delta\ell] + \delta(1 - q)\ell \\ &= qh(1 - \delta) + \delta\ell \end{aligned} \quad [\text{G.7}]$$

Since the seller is strong ($q > \bar{q} = \ell/h$) we see that

$$v(\bar{p}_0) > qh(1 - \delta) + \delta\ell = \ell \quad [\text{G.8}]$$

A strong seller would do better to choose \bar{p}_0 rather than ℓ as the first period price. Hence the only possible candidate for a PBE with $p_0 \in (\ell, \bar{p}_0]$ has the seller setting the first period price \bar{p}_0 and getting a payoff of $qh(1 - \delta) + \delta\ell$.

Now consider sub-games starting with an initial price in the range $(\bar{p}_0, h]$. We will show that there is a constant acceptance probability $0 < a_0(p_0, h) = \bar{a}_0 < 1$ for the type h which yields an equilibrium for this range of sub-games. First, note that if the type h set his acceptance probability for p_0 in this range at $a_0(p_0, h) = 1$ there would not be an equilibrium. If type h was certain to accept such a p_0 the seller would be sure that she faced type ℓ if p_0 was rejected and, since $q_1(p_0) = 0$, [G.5] would lead her to set $p_1 = \ell$.

has terminated, but if it was rejected she can use the additional information to revise her probability beliefs about the type of buyer she faces. We assume that she uses Bayes's Rule (see the Appendix to this chapter) to update her beliefs. We denote S 's updated probability that the buyer is type h given that the period 0 offer was rejected by $q_1 = q_1(p_0)$. The notation $q_1(p_0)$ indicates that the information conveyed by a rejection may vary with the offer rejected.

The equilibrium concept we use is the *perfect Bayesian equilibrium* (PBE), which is a refinement of the sub-game perfect equilibrium used in section F. A pair of strategies is a PBE if they constitute best replies to the other player's strategy in all sub-games, given that in each sub-game each player will use the past behaviour of the other and their knowledge of the other's strategy to update their probability beliefs using Bayes's Rule.

For a PBE the strategies of the players must be equilibrium strategies for any sub-game. The seller, by her initial offer p_0 , can force the game into the sub-game starting with B 's response to that particular initial offer. She chooses which sub-game will be played and will, in equilibrium, choose the sub-game which maximizes her expected discounted payoff. We therefore characterize the PBE by first working out the equilibrium strategies for the sub-games starting with a given p_0 . The equilibrium for the whole game starting with S 's choice of p_0 is then found by looking for the p_0 which maximizes her expected discounted payoff. We will only consider the case in which the seller would be strong in a one-period game: her prior probability belief q exceeds the critical value \bar{q} defined by [G.2]. (Question 3, Exercise 14G examines the weak seller case in which $q < \bar{q}$.)

The type ℓ buyer's optimal strategy is simple: in any period always refuse any price higher than ℓ and accept in any period any offer which is less than or equal to ℓ . (Buyers realize that it cannot be optimal for S to set p_1 less than ℓ in period 1 and so the type ℓ buyer cannot gain by rejecting $p_0 \leq \ell$ in the hope of getting a lower price in period 1.) Thus the type ℓ buyer's decisions rules are

$$a_i(p_i, \ell) = 1 \quad \text{for } p_i \leq \ell, \quad a_i(p_i, \ell) = 0 \quad \text{for } p_i > \ell \quad (i = 0, 1) \quad [\text{G.3}]$$

The type h buyer's strategy in period 1 is also simple: since his is the last move in the game, he just compares S 's offer p_1 with his valuation of the asset and accepts if $p_1 \leq h$:

$$a_1(p_1, h) = 1 \quad \text{for } p_1 \leq h, \quad a_1(p_1, h) = 0 \quad \text{for } p_1 > h \quad [\text{G.4}]$$

In the last period of the game the seller is in essentially the same position as in the single period game *except that her probability beliefs may have been changed by the buyer's rejection of her first period offer*. We can apply the same kind of argument as in the single period model to conclude that S will set $p_1 = \ell$ or $p_1 = h$. We can therefore describe her decision rule in period 1 by the probability

$$\lambda = \Pr[p_1 = \ell]$$

Given her updated probability belief following the buyer's rejection of her first period offer p_0 , the seller's optimal second period decision rule is

$$\begin{aligned} \lambda &= 0 & \text{if } q_1(p_0) > \bar{q} \\ \lambda &= 1 & \text{if } q_1(p_0) < \bar{q} \\ \lambda &\in [0, 1] & \text{if } q_1(p_0) = \bar{q} \end{aligned} \quad [\text{G.5}]$$

We must now find a period 0 acceptance rule for the type h buyer which together with strategies [G.3], [G.4] and [G.5] yields equilibria of the sub-games starting from a first period offer. First, consider the sub-game starting with a first period offer of $p_0 = \ell$. The type h 's optimal response is clearly to accept this offer since he cannot do better by waiting even if the seller's second period strategy [G.5] resulted in $p_1 = \ell$ for sure:

$$h - \ell > \delta(h - \ell)$$

Thus he will set $a_0(\ell, h) = 1$. This acceptance rule and the decision rules [G.3], [G.4], [G.5] generate an equilibrium in the sub-game starting at $p_0 = \ell$. If the seller sets $p_0 = \ell$, the game will end after one period since both types of buyer will accept the offer, and the seller will get an income of ℓ .

Next, consider sub-games starting with S setting $p_0 > \ell$. The type ℓ buyer, following [G.3], will reject this offer whatever he believed about the period 1 offer. If the type h buyer was sure that $p_1 = \ell$, he would accept $p_0 > \ell$ if and only if

$$h - p_0 \geq \delta(h - \ell)$$

or, equivalently, if and only if

$$p_0 \leq \bar{p}_0 \equiv (1 - \delta)h + \delta\ell \quad [\text{G.6}]$$

Over the range of sub-games starting with $p_0 \in (\ell, \bar{p}_0]$ the acceptance rule $a_0(p_0, h) = 1$ and the decision rules [G.3], [G.4] and [G.5] constitute an equilibrium. If the offer $p_0 \leq \bar{p}_0$ is rejected the seller knows that such behaviour is optimal only for the type ℓ buyer and will revise her probability belief to $q_1(p_0) = 0$. Hence, from [G.4], she will set $p_1 = \ell$ and the asset will be bought by the type ℓ in the second period. If the seller faced the type h buyer the offer $p_0 \leq \bar{p}_0$ would be accepted in the first period, since the type h buyer gains nothing by delaying acceptance.

The seller's expected proceeds from choosing an initial price in the range $(\ell, \bar{p}_0]$, is $qp_0 + \delta(1 - q)\ell$ which is increasing in p_0 . (Remember that when the buyer chooses p_0 she must use the prior probability q since she only acquires information from observing the buyer's response to her initial offer.) If the seller's optimal initial offer is in the range $(\ell, \bar{p}_0]$ it must be equal to \bar{p}_0 and will yield her an expected discounted income

$$\begin{aligned} v(\bar{p}_0) &= q\bar{p}_0 + \delta(1 - q)\ell = q[(1 - \delta)h + \delta\ell] + \delta(1 - q)\ell \\ &= qh(1 - \delta) + \delta\ell \end{aligned} \quad [\text{G.7}]$$

Since the seller is strong ($q > \bar{q} = \ell/h$) we see that

$$v(\bar{p}_0) > \bar{q}h(1 - \delta) + \delta\ell = \ell \quad [\text{G.8}]$$

A strong seller would do better to choose \bar{p}_0 rather than ℓ as the first period price. Hence the only possible candidate for a PBE with $p_0 \in (\ell, \bar{p}_0]$ has the seller setting the first period price \bar{p}_0 and getting a payoff of $qh(1 - \delta) + \delta\ell$.

Now consider sub-games starting with an initial price in the range $(\bar{p}_0, h]$. We will show that there is a constant acceptance probability $0 < a_0(p_0, h) = \bar{a}_0 < 1$ for the type h which yields an equilibrium for this range of sub-games. First, note that if the type h set his acceptance probability for p_0 in this range at $a_0(p_0, h) = 1$ there would not be an equilibrium. If type h was certain to accept such a p_0 the seller would be sure that she faced type ℓ if p_0 was rejected and, since $q_1(p_0) = 0$, [G.5] would lead her to set $p_1 = \ell$.

But, from [G.6],

$$h - p_0 < h - \bar{p}_0 = \delta(h - \ell)$$

so that the type h buyer would be better off rejecting the first period offer and waiting to receive the offer ℓ in the next period. Thus $a_0 = 1$ cannot be part of a sub-game equilibrium in this range of p_0 since it is not optimal against the strategy [G.5] given the Bayesian updating by S .

Nor can there be an equilibrium if $a_0 = 0$ for all $p_0 \in (\bar{p}_0, h]$. If the type h refuses the price p_0 then, since both types refuse it, the seller gains no information from having her initial price refused: $q_1(p_0) = q$. Since $q > \bar{q}$ she will set $p_1 = h$ (see [G.5]) and the type h would do better by accepting a first period offer $p_0 < h$.

If the type h follows the strictly mixed strategy $0 < a_0(p_0, h) = \bar{a}_0 < 1$ he would get an expected payoff

$$\begin{aligned} (h - p_0)\bar{a}_0 + \delta[\lambda(h - \ell) + (1 - \lambda)(h - h)](1 - \bar{a}_0) \\ = (h - p_0)\bar{a}_0 + \delta\lambda(h - \ell)(1 - \bar{a}_0) \end{aligned}$$

The marginal value of increase in \bar{a}_0 to him is $(h - p_0) - \delta\lambda(h - \ell)$. If this is positive he will set $\bar{a}_0 = 1$ and if it is negative he will set $\bar{a}_0 = 0$. Hence he is willing to play the strictly mixed strategy only if $h - p_0 - \delta\lambda(h - \ell) = 0$ or

$$\lambda = (h - p_0)/\delta(h - \ell) \quad [\text{G.9}]$$

Now [G.9] implies that when $p_0 \in (\bar{p}_0, h)$, the seller also must be willing to play a strictly mixed strategy $0 < \lambda < 1$ in the second period. From [G.5] this in turn implies that $q_1(p_0) = \bar{q}$ or, using Bayes's Rule

$$\frac{[1 - a_0(p_0, h)]q}{[1 - a_0(p_0, h)]q + [1 - a_0(p_0, \ell)](1 - q)} = \frac{[1 - a_0(p_0, h)]q}{[1 - a_0(p_0, h)]q + (1 - q)} = \bar{q} \quad [\text{G.10}]$$

Hence, if over the range $p_0 \in (\bar{p}_0, h]$ the type h buyer has a first period acceptance probability of

$$a_0 = \bar{a}_0 = (q - \bar{q})/q(1 - \bar{q}) \quad [\text{G.11}]$$

which satisfies [G.10], the seller would be willing to follow a mixed strategy in the second period satisfying [G.9]. Thus the strategies defined by [G.3], [G.4], [G.5], [G.9] and [G.11] (or [G.10]) are an equilibrium for the sub-games starting with $p_0 \in (\bar{p}_0, h]$.

The expected payoff to S from the sub-game equilibria with $p_0 \in (\bar{p}_0, h]$ is

$$q\bar{a}_0 p_0 + \delta\{(1 - \lambda)q(1 - \bar{a}_0)h + \lambda[q(1 - \bar{a}_0) + (1 - q)]\ell\}$$

which is increasing in p_0 . (The reader should check this by substituting for λ from [G.9] and differentiating with respect to p_0 .) Hence if the seller's optimal first period offer is in the range $p_0 \in (\bar{p}_0, h]$ it will be h , yielding an expected discounted income of

$$v(h) = q\bar{a}_0 h + \delta q(1 - \bar{a}_0)h = qh[\bar{a}_0 + \delta(1 - \bar{a}_0)] \quad [\text{G.12}]$$

An initial price \bar{p}_0 is best for S over the range $p_0 \in [\ell, \bar{p}_0]$ and $p_0 = h$ is best over the range $(\bar{p}_0, h]$. Comparing [G.6] and [G.12], we see that the seller's optimal initial price

will depend on the values of the parameters (h, ℓ, q, δ) . Thus there are two types of equilibrium:

1. *High price equilibrium.* S follows the high price strategy h in period 0 and sets $p_1 = h$ in period 1. There is a probability $q(1 - \bar{a}_0)$ that agreement is reached in period 0, a probability $q\bar{a}_0$ that agreement is delayed until period 1 and a probability $(1 - q)$ that there is no agreement at all. Thus adding an extra period to the bargaining game has, in this case, reduced the efficiency of the outcome: although the probability of no trade is the same, there is a positive probability of delay in reaching agreement. Both types of buyer still get a zero surplus and the expected discounted surplus of the seller is reduced because of the positive probability of delay.

The conclusion that the seller is worse off in the two period game is somewhat surprising since she can get the same high price in both periods as she does in the single period game. The explanation is that setting a price of h in both periods is only an equilibrium strategy if the type h buyer follows a strictly mixed strategy in the first period, so that there is positive probability that the type h does not buy until the second period.

2. *Intermediate price equilibrium.* Here S follows the intermediate initial price strategy of $p_0 = \bar{p}_0$ and sets $p_1 = \ell$. Since the type h accepts \bar{p}_0 and the type ℓ rejects it, there is a probability $(1 - q)$ that agreement is not reached until the second period. With this equilibrium there is a positive probability of delay in reaching agreement and the seller makes a concession over time in that the price she sets is lower in the second period than in the first. Thus we have one possible explanation for two commonly observed features of real world bargaining.

The addition of one period to the one-period model makes a considerable difference to the outcome. Remember that we have assumed that the seller is strong, so that in a one-period game she would set a price of h and with probability $(1 - q)$ there would be no trade. Here trade is only delayed if the seller faces a type ℓ buyer. The extra period provides an opportunity for the seller to acquire information. If her first period offer is rejected she revises her probability belief and reduces her offer, since she is now sure that she faces the type ℓ buyer. This opportunity is not available if there is only one period.

The seller is again worse off in the two-period game than in the one-period game, despite the opportunity for acting on information produced in the first period. In the one period model the seller gets qh which exceeds her payoff from the intermediate price strategy [G.7]: $qh > v(\bar{p}_0) = qh(1 - \delta) + \delta\ell$. The seller would do better with two periods if the type h used his acceptance rule from the one-period game with both periods. S would then set a high price in period 0 and reduce the price to ℓ to sell to the type ℓ in period 1 if there was no sale in period 0. However the type h buyer will reject $p_0 = h$ in period 0 if he knew that $p_1 = \ell$. The seller must reduce p_0 to \bar{p}_0 to induce the type h to buy in period 0, so that she can be sure that she faces a type ℓ buyer if her first period offer is refused.

The h type faces a lower price and is made better off by $h - \bar{p}_0$. The ℓ type buyer faces the second period price ℓ and is no worse off by the addition of the extra period. Since the probability of a type h buyer is q , the effect on the expected combined surplus of buyers

and seller of adding an extra period in the intermediate price case is

$$(h - \bar{p}_0)q + [q\bar{p}_0 + \delta(1 - q)\ell] - hq = \delta(1 - q)\ell > 0$$

Thus the total expected gain from trade is increased, in contrast with the high price equilibrium case.

The values of the parameters (q, h, ℓ, δ) will determine which equilibrium yields the seller the highest expected discount payoff and therefore which one is chosen. For example, if q is close to 1 then $v(h) > v(\bar{p}_0)$. As [G.11] indicates, the equilibrium probability that the type h buyer accepts a first period price of h tends to 1 as q tends to 1, so that $v(h)$ tends to h . Although $v(\bar{p}_0)$ also increases with q it can never exceed $h(1 - \delta)$.

The intermediate price equilibrium will occur if the discount factor δ is small. A small δ makes the intermediate price strategy more attractive for the seller because it increases the intermediate first period price \bar{p}_0 and reduces the significance of the fact that the seller will set a low price in the second period. $v(\bar{p}_0)$ tends to qh as δ tends to 0, whereas $v(h)$ tends to $qh\bar{a}_0$.

Conclusion

The non-cooperative game studied in this section provides an example of how rational individuals may fail to reach agreements or do so only after a delay, despite the fact that their joint gains are maximized by immediate agreement. The model gives a rigorous demonstration of the way in which differences in the information available to the parties can affect both the likelihood of trade and the terms on which trade takes place, a point to which we will return in Chapter 22. However, the non-cooperative approach is still far from providing a complete and useful theory of bargaining, as its rather poor predictive performance seems to indicate (Ochs and Roth, 1989).

As we saw, a full account of the equilibrium of the simple two-period game in this section is surprisingly complex because actions can convey information. A further consequence of actions providing information to the less well-informed party is that predictions of models with incomplete information are often strongly dependent on fine, and apparently arbitrary details of the bargaining process. Details, such as whether the parties make simultaneous or sequential moves, or whether one or both can make offers, matter because they affect the way in which actions transmit information. The difference in the results between the one and two-period bargaining model in this section arise because in a two-period model the action of the better informed party (acceptance or rejection of the first period offer) conveys information to the less well-informed party. There are equally dramatic consequences if it is assumed that the better informed party is the one who makes the offers (see Question 6, Exercise 14G). Unfortunately, it is often essentially arbitrary to adopt one set of assumptions about the bargaining process, and therefore the structure of the game, rather than another. For example, in a model of union-firm bargaining it is plausible that the firm is better informed about its revenue from employing union members than the union (see Question 4, Exercise 14G). We should therefore reflect this in the specification of what the parties know at the start of the game. But there is no obvious reason why we should assume that it is the union, or the firm, which makes the offers to which the other

party responds. The predictions of the model will, however, depend on the identity of the party making the offer, since offers by the firm may reveal its information about demand conditions to the union. Until we can resolve such issues the theory of non-cooperative bargaining will consist of a set of interesting but essentially *ad hoc* models.

Exercise 14G

1. *Continuum of buyer types.* Suppose that in the one period game, b is uniformly distributed over the interval $[\ell, h]$. What price will S propose and what is the probability that trade will take place? What is the expected gain from trade and how is it split between S and B ?
2. *Efficiency.* Is the outcome of bargaining in the single period model inefficient in the sense that there exists an alternative allocation mechanism which could make seller and buyers better off and which does not require a regulator with better information than the seller?
3. *Equilibrium with weak seller.* Show that there are two possible PBE for the two period game with a weak seller ($q < \bar{q}$), one of which results in the seller setting $p_0 = \ell$ and the other in the seller choosing $p_0 = \bar{p}_0$ and $p_1 = \ell$.
4. *Union-firm bargaining with asymmetric information.* Use the results of this section to examine the outcome of union and firm bargaining when the union makes the offers but only knows that the firm's revenue function is either $hR(f(z))$ with probability q or $\ell R(f(z))$ with probability $(1 - q)$.
5. *Commitment.* Suppose that S could commit herself at period 0 to the prices she offers in period 0 and period 1. What prices will she set?
6. *Offers by the informed player.* In the following variations on the game in the text both players make offers. Derive and compare the PBE. (a) In period 0 the seller makes an offer and the buyer accepts or rejects. If B rejects then in period 1 he makes an offer and S accepts or rejects. (b) In period 0 B makes an offer and S accepts or rejects. If S rejects p_0 she makes a proposal p_1 which B may accept or reject. (Hint: remember that B 's action in period 0 may convey information.)

Appendix: Bayes's Rule

Consider a random variable x which takes on the values x_1, x_2, \dots, x_n . For example, x may be (a) the amount of rainfall in an area in a given year; (b) the value a buyer attaches to an asset; (c) a dichotomous variable taking on the value 0 when the individual is a good and 1 when he is a bad driver. Before any other information is received the *prior* probability that x has the value x_i is $P(x_i)$.

Suppose that a signal s about the value of the random variable x is received. The signal s received varies with the random variable of interest but also with other random and unobservable factors. The signal can take on the values s_1, s_2, \dots, s_m . In the previous

examples s may be (a) the hours of sunshine in the area in the year; (b) the event that a buyer accepts or refuses an offer of a given size. In the case (c), where the random variable is driving skill, the signal could be the number of accidents in a given period. We denote the conditional probability of receiving signal s_j when x has value x_i by $P(s_j|x_i)$. From the laws of probability

$$P(s_j|x_i) = \frac{P(s_j \text{ and } x_i)}{P(x_i)} \quad [1]$$

The conditional probability that the random variable has the value x_i given that the signal s_j is received is

$$P(x_i|s_j) = \frac{P(x_i \text{ and } s_j)}{P(s_j)} \quad [2]$$

But from [1] we can substitute $P(s_j|x_i)P(x_i)$ for $P(x_i \text{ and } s_j) = P(s_j \text{ and } x_i)$ to get

$$P(x_i|s_j) = \frac{P(s_j|x_i)P(x_i)}{P(s_j)} \quad [3]$$

which is *Bayes's Theorem*.

Bayes's Theorem shows how new information (the value of the signal) can be used to produce a *posterior* probability distribution from the *prior* probabilities $P(x_i)$ and the *relative likelihood ratio* $P(s_j|x_i)/P(s_j)$. For example, suppose that: x_i is a high level of rainfall; s_j is a large number of hours of sunshine; high sunshine is less likely when there is high rainfall [$P(s_j|x_i) < P(s_j)$]. Then the signal s_j would lead us to place a lower probability on the occurrence of high rainfall. The posterior probability of high rainfall, given high sunshine, would be smaller than the prior probability: $P(x_i|s_j) < P(x_i)$.

The posterior probability is often written in a more complicated form by using the fact that $P(s_j) = \sum_k P(s_j \text{ and } x_k) = \sum_k P(s_j|x_k)P(x_k)$ to get

$$P(x_i|s_j) = \frac{P(s_j|x_i)P(x_i)}{\sum_k P(s_j|x_k)P(x_k)} \quad [4]$$

This is the form used in [G.10], where x_i denotes a high type buyer, the signal s_j is rejection of the offer p_0 , so that $P(x_i) = q$, $P(s_j|x_i) = 1 - a_0(p_0, h)$ and $P(x_i|s_j) = q_1(p_0)$.

References and further reading

There is an extensive discussion of the demand for inputs by the competitive firm in:

- C. E. Ferguson. *The Neo-classical Theory of Production and Distribution*, Cambridge University Press, London, 1969, chs 6, 8, 9.
J. Hicks. *Theory of Wages*, Macmillan, London, 1964.

and the cost function is used to examine the competitive market input demand in:

- W. E. Diewert. 'A note on the elasticity of derived demand in the n -factor case', *Economica*, May 1971, 192-8.

On the behaviour and objectives of unions see:

- H. Farber. 'The analysis of union behavior,' in *Handbook of Labor Economics*, Vol. 1, A. Ashenfelter and R. Layard (eds), Elsevier, Amsterdam, 1986.

The material in sections E, F and G is based on:

- J. F. Nash. 'The bargaining problem', *Econometrica*, 18, 1950, 155-62.
A. Rubinstein. 'Perfect equilibrium in a bargaining model', *Econometrica*, 50, 1982, 97-109.
D. Fudenberg and J. Tirole. 'Sequential bargaining with incomplete information', *Review of Economic Studies*, 50, 1983, 221-47.

The Nash paper is simple enough to serve as an introduction to the Nash bargaining solution but the others are best approached after the following:

- K. Binmore and P. Dasgupta (eds). *The Economics of Bargaining*, Basil Blackwell, Oxford, 1987.
M. J. Osborne and A. Rubinstein. *Bargaining and Markets*, Academic Press, San Diego, 1990, chs 1-5.

For some experimental tests of alternative bargaining models see:

- J. Ochs and A. E. Roth. 'An experimental study of sequential bargaining', *American Economic Review*, 79, 1989, 355-84.

There is a survey of disagreement and delay in union-firm negotiations in:

- J. Kennan. 'The economics of strikes', in *Handbook of Labor Economics*, ch. 19, Vol. 2, A. Ashenfelter and R. Layard (eds), Elsevier, Amsterdam, 1986.

CHAPTER 15

Investment and consumption over time

A. Introduction

The theory of the consumer developed in Chapter 3 related to choice of consumption goods in a single time period. It took no account of saving and dissaving, or lending and borrowing, which we would normally expect to be an important aspect of consumer behaviour. Moreover, we know that the operation of the capital market – the market for borrowing and lending – influences the economy in important ways, and so it is useful to develop a theory of the operation of that market.

In the theory of the firm, the process of change in equilibrium scale in the long run can be viewed in a different way from the approach in Chapter 8. The firm changes its scale by investing in new capacity, and so we could view the problem of determining long-run equilibrium output as the problem of choosing the most profitable amount of investment to produce it. Hence it is instructive to construct a theory of how such investment decisions are taken. Section B examines the problem of optimal intertemporal consumption, section C the investment decision and section D the capital market. Section E shows how the two-period model of previous sections may be to many periods and to allow for adjustment costs and depreciation.

B. Optimal consumption over time

We begin with the theory of the consumer. Assume that time is divided into equal discrete intervals – say into years. Given the consumer's annual income, we assume that the atemporal theory applies, and the consumer allocates this income optimally over goods. However, he also has a further choice, which is either to lend some of his income, or to borrow. Lending will imply a reduction in current consumption, but an increase in future consumption, and conversely for borrowing. Thus, the analysis of borrowing and lending decisions is the analysis of the consumer's choice of a consumption pattern over time.

In analysing these decisions we find it convenient to make the following assumptions:

- Within each period of time prices are given and the consumer spends his consumption budget optimally, which implies that we can conduct the analysis entirely in terms of choices of values of the *total sum* of consumption expenditure in each period, rather than of quantities of particular goods and services. (Recall the composite commodity theorem of section 4D)
- To reduce everything to two dimensions, we assume that only two time periods exist, time 0 (the present) and time 1 and denote the individual's total consumption expenditures in the two periods by M_0 and M_1 respectively.
- The consumer faces a perfectly competitive capital market so that there is a given price for borrowing and lending, which is usually expressed as an interest rate, r . Thus £100 borrowed at time 0 implies that £100(1 + r) must be repaid at time 1 (r is therefore defined as an *annual* interest rate), so $1 + r$ can be thought of as the price paid for borrowing, or received for lending £1. Since the capital market is perfect, all borrowers and lenders regard themselves as being able to borrow or lend as much as they like at the going rate of interest r .

Given these assumptions, we can construct a model of choice of consumption over time. The elements of the consumer's optimization problem are:

- The choice variables M_0 and M_1 .
- We assume that the consumer has a preference ordering over combinations of current and future consumption expenditure (M_0, M_1), and that this ordering satisfies all the assumptions made in section 3A. Hence preferences can be represented by the utility function $u(M_0, M_1)$ and indifference curves have the usual shape. Since the consumer prefers more expenditure in one period to less, other things being equal, the marginal utility of current and future consumption is positive.
- The feasible set is defined as follows. The consumer will be initially endowed with an income time-stream (\bar{M}_0, \bar{M}_1) , $\bar{M}_0, \bar{M}_1 \geq 0$. So that the problem is not trivial, we assume at least one of these is positive. Let A represent the amount the consumer borrows or lends in year 0, with $A > 0$ for borrowing, and $A < 0$ for lending. Then, the consumer's feasible consumptions will be constrained by:

$$0 \leq M_0 \leq \bar{M}_0 + A \quad [\text{B.1}]$$

$$0 \leq M_1 \leq \bar{M}_1 - (1 + r)A \quad [\text{B.2}]$$

The right-hand inequalities bind in [B.1] and [B.2] as a result of our non-satiation axiom. The optimal point will be on a boundary of the feasible set. Then, solving for A in [B.2] and substituting into [B.1] gives:

$$(M_0 - \bar{M}_0) + \frac{(M_1 - \bar{M}_1)}{1 + r} = 0 \quad [\text{B.3}]$$

or:

$$M_0 + \frac{M_1}{1+r} = \bar{M}_0 + \frac{\bar{M}_1}{1+r} = V_0 \quad [\text{B.4}]$$

Equation [B.3] should be compared with the consumer's budget constraint in section 3E. Clearly, the present case is directly analogous with that analysed there, $(M_0 - \bar{M}_0)$ and $(M_1 - \bar{M}_1)$ representing net demands for consumption in the two periods, and $1/(1+r)$ the relative price. Equation [B.4] is the consumer's *wealth constraint*. The value of the consumer's chosen consumption time-stream is equal to the value of his endowed income time-stream. These values are expressed in terms of income at time 0, i.e. they are *present values*. V_0 is the present value of the consumer's endowed income time-stream, or in other words his wealth. The economic interpretation of [B.4] is as follows: by borrowing or lending the consumer may achieve a consumption time-stream which differs from his endowed time-stream, but in doing this he is constrained by his wealth.

The assumption of a perfect capital market implies that r can be taken as constant by the consumer, and so the wealth constraint can be graphed as a straight line, known as a *wealth line*, such as V_0 in Fig. 15.1. The slope of this line is:

$$\frac{dM_1}{dM_0} = -(1+r) \quad [\text{B.5}]$$

since [B.4] implies the equation $M_1 = (1+r)V_0 - (1+r)M_0$. Note that it *must* pass through the initial endowment point $\bar{M} = (\bar{M}_0, \bar{M}_1)$, since this point always satisfies [B.4]. The wealth line V_0 represents the set of market exchange opportunities to the consumer; by lending, he can move leftwards from \bar{M} along the line; by borrowing, he moves rightward. Each point on V_0 represents simultaneously a consumption time-stream, an amount of borrowing or lending in year 0, and a corresponding repayment in year 1.

The absolute value of the slope of V_0 is determined by r . A reduction in r leads to a flatter line, such as V'_0 in the figure. Note that this line must continue to pass through \bar{M} .

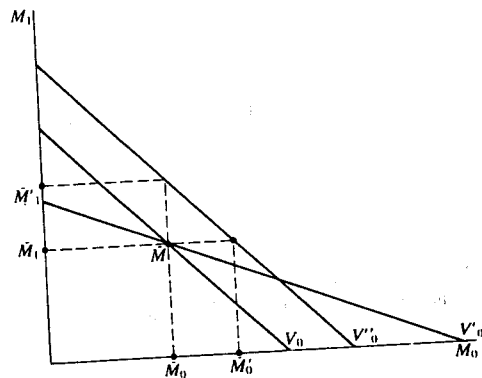


Fig. 15.1

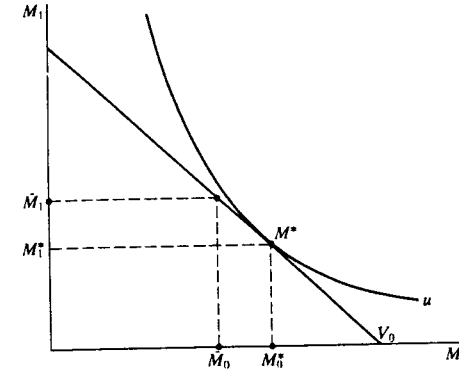


Fig. 15.2

since the initial endowment point continues to satisfy [B.4]. Hence, changes in r cause the line to rotate through \bar{M} . A change in an initial endowment, r remaining unchanged will change the intercept of V_0 but not the slope of the line, and so the line shifts parallel to itself, for example to V''_0 . This line corresponds to, say, an increase in \bar{M}_0 to \bar{M}'_0 , or an increase in \bar{M}_1 to \bar{M}'_1 , or intermediate increases in both. The new wealth line must pass through the new initial endowment point.

Consider the solution to the consumer's optimization problem. Given our assumptions about preferences, we expect to obtain a tangency solution such as that in Fig. 15.2. (Why is it reasonable to assume that we would not have a corner solution in this case?) As drawn, we have a tangency solution at M^* , with the consumer's chosen consumption time-stream at (M^*_0, M^*_1) . This is achieved by the consumer's borrowing an amount $M^*_0 - \bar{M}_0$ in year 0, and having to repay $\bar{M}_1 - M^*_1 = (1+r)(M^*_0 - \bar{M}_0)$ in year 1. It would be possible for the indifference curves, initial endowments, or interest rate to be such that the optimal point implies lending (M^* to the left of \bar{M} on V_0), or neither borrowing nor lending (M^* coincides with \bar{M}), and these cases are left to the reader.

At the optimum the slope of the indifference curve equals the slope of the wealth line. But the slope of the indifference curve is the negative of the ratio of marginal utilities, $-u_0/u_1$, where u_i is the marginal utility of period i consumption (see Chapter 3, section A). At the optimum, therefore, using [B.5],

$$\frac{u_0}{u_1} = 1+r \quad [\text{B.6}]$$

Now a £1 reduction in M_0 reduces u by the marginal utility u_0 . There will exist an increase in M_1 which will make the consumer just as well off as before the £1 reduction in M_0 . This compensating increase in M_1 is £ $(1+r)$ and is defined by

$$u_0(M_0, M_1) \equiv u_1(M_0, M_1)(1+r) \quad [\text{B.7}]$$

where the notation emphasizes that the marginal utilities u_0 and u_1 depend on the consumption time-stream. In words, raising M_1 by £1 increases u by the marginal utility

u_1 so that increasing M_1 by $\mathcal{L}(1 + \rho)$ will raise u by $u_1(1 + \rho)$ and this just offsets the effect of the $\mathcal{L}1$ reduction in M_0 . Note that since $u_0 > 0$ and $u_1 > 0$ we must have $1 + \rho > 0$ so that $\rho > -1$. ρ can be interpreted as the consumer's *subjective rate of interest* since it shows how much *extra* consumption in period 1 is required to compensate for the loss of $\mathcal{L}1$ of current consumption. ρ may be negative if less than $\mathcal{L}1$ extra of M_1 is required. It is subjective because it depends on the consumer's preferences, not on observable market phenomena. Since u_0 and u_1 will depend on M_0 and M_1 so must ρ : $\rho = \rho(M_0, M_1)$. Rearranging [B.7], we get

$$\frac{u_0}{u_1} \equiv 1 + \rho \quad [\text{B.8}]$$

and we see that as the consumer moves along an indifference curve from left to right substituting M_0 for M_1 , ρ will decline since the slope of the indifference curve becomes flatter, and current consumption relatively less valuable. ρ is also known as the consumer's *rate of time preference*. We can use [B.8] to write the optimum condition in [B.6] simply as

$$\rho = \rho(M_0, M_1) = r \quad [\text{B.9}]$$

and this (or [B.6]) together with the constraint [B.4] provides two equations to determine the optimal M_0^* , M_1^* . The consumer is in equilibrium where his subjective rate of interest is equal to the market rate of interest. In other words, the consumer lends up to the point at which the market interest rate is just sufficient to compensate for the marginal reduction in current consumption. Alternatively, he borrows until he reaches the point at which the price he must pay (in terms of reduced consumption next period) is just sufficient to offset the value to him of the additional consumption this period.

Exercise 15B

- Explain in commonsense terms the inequalities in [B.1] and [B.2].
 - Draw diagrams analogous to Fig. 15.2, in which (i) the consumer lends at the optimum, and (ii) he neither borrows nor lends.
 - Explain why a corner solution (with $M_0^* = 0$ or $M_1^* = 0$) would be intuitively unreasonable.
 - Explain why a constant time preference rate would be intuitively unreasonable.
 - What would be implied by values of ρ equal, respectively, to -0.2 , 0 and 0.2 ?
- Derive the demand functions for present (period 0) and future (period 1) consumption as functions of wealth and the rate of interest, for a consumer who has the Cobb-Douglas utility function $u = M_0^\alpha M_1^{1-\alpha}$ ($0 < \alpha < 1$).
- Discount factor.** The consumer's subjective discount factor δ is the rate at which the consumer is willing to give up period 0 consumption for period 1 consumption: $\delta \equiv u_1/u_0$. Is it always the case that $\delta < 1$? Recast the description of the consumer's optimum in terms of the discount factor.
- Imperfect capital market.** Suppose, because of transaction costs or taxation of interest income that the interest rate at which the consumer can borrow exceeds that at which he can lend (though both are still invariant with the quantity borrowed or

lent). Construct the feasible set in this case, and suggest the solution possibilities. Give, in terms of interest rates and the consumer's time preference rate, the condition which holds at the optimal solution for a consumer who neither borrows nor lends.

C. The optimal investment decision

A 'firm' can be regarded for the moment as a single decision-taker who has available some specific set of *productive investment opportunities*, i.e. some means of transforming current income into future income, by means of production rather than exchange. The investment and production decisions taken by the owner of the firm will determine the cash flows he receives from the firm in each period. At the same time the owner has access to the capital market on which she can borrow or lend, and so her consumption expenditure in each period need not equal the cash flow generated by her investment and production decisions. (We assume that the firm is the only source of income for the owner.) The owner of a firm with investment opportunities must therefore solve both the *investment and production decision problem*, which determines the firm's cash flow, and the *consumption decision problem*, which determines how her consumption expenditures differ from the income she receives from the firm.

We proceed by making the same two-period and perfect capital market assumptions as we did in the previous section in the case of the consumer. We also assume that the owner of the firm has a preference ordering over the consumption time-streams (M_0, M_1) and that we can represent this by indifference curves in the usual way. The only new element in the analysis is the feasible set, which now depends on the technological possibilities of production and investment as well as the terms on which the owner can borrow or lend in the capital market.

Production and investment possibilities

The cash flow or dividend D_t ($t = 0, 1$) received from the firm by the owner in each period is the firm's revenue less its expenditure. If labour is the only variable input, the expenditure in each period is the sum of the amount paid in wages and the outlay on purchases of additional physical capital, i.e. investment. Hence the cash flow is

$$D_0 = pf(L_0, K_0) - wL_0 - p_K(K_1 - K_0) = pf(L_0, K_0) - wL_0 - I \quad [\text{C.1}]$$

$$D_1 = pf(L_1, K_1) - wL_1 \quad [\text{C.2}]$$

where p , w and p_K are the competitive market prices of the firm's output, labour and physical capital; $f(L_t, K_t)$ is the firm's production function, L_t and K_t the labour and capital inputs, in period t . We assume for simplicity that all prices and the form of the production function are constant over time. K_0 is the stock of physical capital inherited at the start of period 0 and $I = p_K(K_1 - K_0)$ is the expenditure by the firm in period 0, for the purpose of increasing its capital stock to K_1 for use in production in period 1. Again for simplicity, we assume that there is no depreciation of the capital stock. Note

that there will be no investment in the second period since there is no third period, and also that if $K_1 < K_0$ the firm is disinvesting, i.e. selling off some of its capital to increase its cash flow in the first period (section E contains more general formulations, with many periods, depreciation and adjustment costs.)

In earlier chapters we defined the input variables in the production function as *flows* per period. Here we have a production function in which one of the inputs (L_t) is also a flow per period (so many hours of labour supplied by the workforce) but the other (K_t) is a stock (so many machines, say). The production function used here is compatible with our earlier approach because we can let $k_t = H(K_t)$ be the flow of capital services from a capital stock of size K_t . Then if the production function relating the output per period to the flow of labour and capital services per period is $y_t = \hat{f}(L_t, k_t)$, the production function used here is just $y_t = f(L_t, K_t) \equiv \hat{f}(L_t, H(K_t))$.

The firm will always choose the labour input to maximize the cash flow in each period for given levels of the capital stock, since choice of L_t affects only the cash flow of that period and the owner will always prefer a larger cash flow to a smaller in a period given that the other period's cash flow is unaffected. Given the optimal choice of the variable input L_t in each period and the fixed initial capital stock K_0 , the cash flows in each period depend only on the capital stock K_1 chosen for period 1:

$$D_0 = pf(L_0^*, K_0) - wL_0^* - p_K(K_1 - K_0) = D_0(K_1) \quad [C.3]$$

$$D_1 = pf(L_1^*, K_1) - wL_1^* = D_1(K_1) \quad [C.4]$$

where L_t^* is the optimal level of the labour input in period t . Since both periods' cash flows depend on K_1 we can derive a relationship between D_0 and D_1 by varying K_1 , i.e. by investing (or disinvesting). D_0^* does not depend on K_1 , since the marginal cash flow from variations in L_0 is $p \partial f(L_0, K_0) / \partial L_0 - w$, which does not vary with K_1 . Hence increasing K_1 reduces D_0 at the rate p_K . L_1^* will, however, depend on K_1 since the marginal period 1 cash flow from variations in L_1 is $p \partial f(L_1, K_1) / \partial L_1 - w$, which is affected by changes in K_1 . Hence

$$\frac{dD_1}{dK_1} = p \frac{\partial f(L_1^*, K_1)}{\partial K_1} + \left(p \frac{\partial f(L_1^*, K_1)}{\partial L_1} - w \right) \frac{dL_1^*}{dK_1} = p \frac{\partial f(L_1^*, K_1)}{\partial K_1} \quad [C.5]$$

where we have used the fact that L_1^* maximizes D_1 for given K_1 and so the marginal period 1 cash flow from L_1 is zero at L_1^* . (This is yet another example of the Envelope Theorem of section 2J, chapter 2.) Increasing K_1 will therefore reduce D_0 and increase D_1 (as long as the marginal product of capital is positive in period 1).

Fig. 15.3 plots feasible combinations of cash flows that the owner of the firm can receive by varying her investment decision, i.e. by altering K_1 and so moving along the curve PP . $\bar{D} = (\bar{D}_0, \bar{D}_1)$ is assumed to be the cash flow time-stream the firm will generate if it neither invests nor disinvests, so that $K_1 = K_0$ and $I = 0$. By increasing K_1 say to K_1^* through investing $I' = p_K(K_1^* - K_0)$, the cash flow of the first period is reduced to $D_0' = D_0 - I'$ and the next period's cash flow increased to D_1' . (The exercises at the end of this section (15C) ask you to investigate influences of the production function and p , w and p_K on the shape of curve PP .)

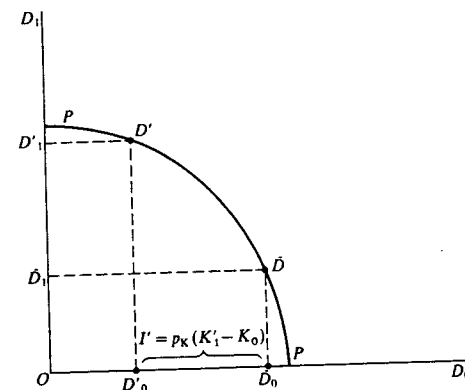


Fig. 15.3

Borrowing and lending possibilities

The owner of the firm has access to a capital market on which she can borrow or lend at the interest rate r and so her consumption expenditure time-stream (M_0, M_1) may differ from the cash flow time stream (D_0, D_1) she receives from the firm.

Fig. 15.4 combines wealth lines, similar to those of section B, with the curve PP from Fig. 15.3. It shows the feasible set when there are possibilities of altering the time pattern of consumption by both production and capital market activities, i.e. by moving along PP and along a wealth line (borrowing or lending).

By investing or disinvesting the owner of the firm can move along PP and achieve different combinations of cash flows. Given her investment decision (choice of D_0, D_1) the owner can then enter the capital market and trade (borrow or lend) to any point along the wealth line through her cash flow combination. For example if the owner does not invest or disinvest she will be at the point \bar{D} on PP . She could then lend out some of her period 0 cash flow and move along the wealth line \bar{V} to a point such as M' where she is better off (on a higher indifference curve than \bar{u} through \bar{D}).

Higher wealth lines will present the owner with better consumption possibilities than lower wealth lines since there will always be some point on the higher wealth line which is on a higher indifference curve than *all* the points on the lower wealth line. Hence the firm's optimal investment decision (choice of K_1 or equivalently choice of D_0, D_1) will be that which maximizes the owner's wealth. In Fig. 15.4 this is the cash flow combination D_0^*, D_1^* achieved by choosing a second period capital stock of K_1^* and investing $I^* = p_K(K_1^* - K_0)$ in the first period. V^* is the highest possible wealth line attainable by investment along PP and so by investing I^* the owner is put in the best possible position for engaging in borrowing or lending in the capital market.

Given the optimal wealth maximizing investment and production decisions the owner chooses some combination of consumption expenditures in the two periods (M_0, M_1) along the maximum wealth line V^* .

Although the solution to the investment decision problem was independent of these preferences, it presupposed that the optimal finance decision would also be taken. In the case in which the firm is owned by two or more shareholders, while decisions are taken by a salaried manager, it may be thought that two kinds of problems would arise. First, the preferences of the shareholders are likely to differ, and so there appears to be the possibility of conflict. Second, how does the manager, to whom choice has been delegated, obtain information on shareholders' preferences, so as to make decisions in accordance with them?

We now establish an important proposition which shows that, if shareholders are able to borrow and lend on the capital market, neither of these problems arises. The proposition is: *if the capital market is perfect, then the manager of the firm acts in the best interests of the shareholders by choosing investment so as to maximize the present value of the firm's income stream.* We prove this proposition as follows.

The i th shareholder in the firm owns a proportion of the issued share capital, s_i , $i = 1, 2, \dots, n$, which entitles her to receive the share s_i in the income of the firm. Each shareholder will wish to choose a consumption time-stream which maximizes her utility, subject to her wealth constraint. But her endowed wealth depends at least in part on the income stream which will accrue from ownership share in the firm, i.e.

$$V_i = V_a + s_i D_0 + s_i \frac{D_1}{1+r} = V_a + s_i V$$

where V_i is the i th shareholder's wealth, V_a that arising from sources other than the firm in question, and $V = D_0 + [D_1/(1+r)]$ the present value of the firm's income stream. Clearly, given V_a , maximizing V is in the interest of every shareholder, regardless of her particular preferences, since it puts her on the highest possible wealth constraint. Thus the firm's manager can choose optimal investment as before, and the only information he requires is the value of the market interest rate r . It is the *consumption decision* which changes for the shareholder-owned firm. There is now no single optimal solution to this, given $n(>1)$ shareholders with differing preferences. The most straightforward solution is for the firm to distribute to shareholders the net income stream resulting from the optimal investment choice, and shareholders are then able to adopt their own borrowing or lending policies in such a way as to attain their overall optimal positions.

We can view this result as an implication of the separation between investment and consumption decisions described earlier. If the optimal investment decision can be taken independently of preferences, then the existence of a number of shareholders with diverse preferences does not create problems. More light is cast on the role of the perfect capital market assumption if we adopt a slightly different interpretation of the proposition. Since all the shareholders can borrow or lend on the capital market at the same rate of interest, then each will be in equilibrium at the point at which her time preference rate equals the market interest rate. However diverse the general structure of their preferences therefore, each values a marginal increment of future income at the same rate in terms of current income, i.e. at the rate measured by the market interest rate r . In taking investment decisions on behalf of shareholders, the managers of the firm can use this interest rate to evaluate gains in future income against sacrifices in current consumption. When the capital market

is imperfect or incomplete, shareholders will disagree about the best investment policy because they will place different values on the benefits from the investment – the changes in their income streams. We return to this question in chapter 22, section H.

Present value and profit maximizing models

In this section we have outlined a model of the firm which chooses its variable input in each period and its capital stock for the next period so as to maximize the *present value of its cash flow*. In Chapter 9 we developed a model of the firm based on the assumption of *profit maximization*. We will now show that these two models are in fact equivalent.

The problem of maximizing the present value of the firm's cash flow is

$$\max_{L_0, L_1, K_1} V = D_0 + \frac{D_1}{(1+r)} \quad [\text{C.10}]$$

where D_0 and D_1 are defined as in [C.1] and [C.2]. The first-order conditions are

$$\frac{\partial V}{\partial L_0} = \frac{\partial D_0}{\partial L_0} = p_0 f_L - w = 0 \quad [\text{C.11}]$$

$$\frac{\partial V}{\partial L_1} = \frac{\partial D_1}{\partial L_1} \cdot \frac{1}{(1+r)} = \frac{1}{(1+r)} (p_1 f_L - w) = 0 \quad [\text{C.12}]$$

$$\frac{\partial V}{\partial K_1} = \frac{\partial D_0}{\partial K_1} + \frac{\partial D_1}{\partial K_1} \cdot \frac{1}{(1+r)} = -p_K + \frac{1}{(1+r)} p_1 f_K = 0 \quad [\text{C.13}]$$

where f_L, f_K are the marginal products of labour and capital and p_t is the price of output in period t and p_t may differ in the two periods. We see that the variable labour input is chosen in each period so that $p_t = w/f_L$. Since (from Chapter 8, section C) w/f_L is short-run marginal cost, we have shown that, just like the profit maximizing firm of Chapter 9, section B, the present value maximizing firm will always choose a variable input level (and hence output) where price is equal to short-run marginal cost.

From [C.13] the firm's investment decision (choice of K_1) satisfies

$$p_1 = \frac{p_K}{f_K} (1+r) = \frac{w}{f_L} \quad [\text{C.14}]$$

An additional unit of capital bought in period 0 for use in period 1 costs p_K in cash flow foregone in period 0. Given the market rate of interest r , p_K foregone in period 0 is equivalent to $p_K(1+r)$ foregone in period 1, since a loan of p_K in period 0 would have to be repaid at a cost of $p_K(1+r)$ in period 1.

$p_K(1+r)$ is the opportunity cost of an additional unit of physical capital in terms of period 1 cash flow. The middle term in [C.14] is the marginal cost in terms of period 1 cash flow of output in period 1 produced by installing more capital paid for in period 0. w/f_L is the marginal cost in terms of period 1 cash flow of producing output by hiring more labour (which must be paid for in that period). [C.14] expresses the requirement that when both inputs are variable, i.e. in the *long run*, the firm plans to produce where

long-run marginal cost equals price and it chooses its fixed input on this basis. Actual output in period 1 is chosen by varying the labour input in period 1 so that short-run marginal cost is equal to price. If, as in this case, the firm accurately forecasts p_1 , long-run marginal cost will also turn out to be equal to price. Thus maximizing the present value of the firm's cash flow and equating long-run marginal cost to price are equivalent formulations of the firm's investment decision.

We saw in Chapter 10 that for a competitive industry to be in long-run equilibrium every firm must have chosen the level of its fixed input so that price equals long-run marginal cost (so that it has no wish to alter its plant size) and every firm should just be breaking even, i.e. price equals long-run average cost (so that no firm wishes to enter or leave the industry). We can restate these requirements in the equivalent terms of the present value-maximizing firm's investment decision.

Present value maximization requires that K_1 be chosen so that [C.14] holds and [C.14] can be rearranged to yield

$$r = \frac{p_1 f_K}{p_K} - 1 = \frac{p_1 f_K - p_K}{p_K} \quad [\text{C.15}]$$

But the right-hand side of [C.15] is merely i , the marginal rate of return. (To see this use [C.1] and [C.2] to substitute for D_0 and D_1 in [C.6] which defines i .) Hence, as we noted in discussion of [C.8], optimal investment implies that $r = i$ and this is the equivalent of the condition that price equals long-run marginal cost.

There will be no incentive for firms to enter the industry if the present value of the cash flows generated by starting production is non-positive. Since it takes one period to install new plant the present value of the cash flow from entry is

$$\frac{D_1}{(1+r)} - p_K K_1 \quad [\text{C.16}]$$

since the new firm will not be producing any output in period 0. [C.16] is also the present value of the *additional* cash flow to an existing firm which decides to continue in production, i.e. to choose a positive K_1 . [C.16] must therefore be non-negative for a firm which chooses to stay in the industry. Hence [C.16] must be zero if there is to be neither entry nor exit from the industry or

$$r = \frac{D_1 - p_K K_1}{p_K K_1} \quad [\text{C.17}]$$

Now $D_1 - p_K K_1$ is the return from investing $p_K K_1$ in the industry so that $(D_1 - p_K K_1)/p_K K_1$ is the average rate of return, g . Hence if the industry is in long-run equilibrium the average rate of return being earned in the industry will be equal to the market rate of interest. This is an equivalent formulation of the price equals long-run average cost condition for long-run equilibrium in a competitive industry.

Thus in long-run competitive equilibrium the rate of return on capital to *all* firms is equal to the market interest rate. If $g > r$ then capital is moved into the industry in search of an excess profit, while if $g < r$ capital is moved out because a higher profit can be earned elsewhere, e.g. by lending on the market. Thus the equalization of gross profit rates with each other and with the market rate of interest is a long-run tendency of a competitive

economy. This conclusion must be modified if there are different degrees of risk in different industries and owners are averse to risk. Riskier industries would have to have a greater rate of return to compensate for greater risk but we would still predict that in a competitive economy industries in similar risk categories would have the same profit rates. We take up the analysis of risk and uncertainty in the last four chapters.

Exercise 15C

- Examine the effects on the curve PP in Fig. 15.3 of
 - changes in the price of output, the capital good and labour;
 - the capital good depreciating at a constant percentage rate per period;
 - disinvestment being impossible.
- How will the firm's investment and production decisions be affected by the changes in PP due to the factors listed in Question 1?
- Show in Fig. 15.4 cases in which:
 - no investment would be undertaken;
 - D_1 would be maximized;
 and state necessary and sufficient conditions for each of these.
- Show that at an optimal solution to the firm's problem, $i = r = \rho$: where ρ is the owner's rate of time preference.
- Show that as long as the shareholder-owned firm makes the optimal investment choice (maximizes $D_0 + D_1/(1+r)$), any borrowing or lending policy it then adopts leaves the shareholders' utility unaffected (provided of course that it distributes to shareholders all income flows resulting from investment, borrowing or lending). (*Hint*: consider the borrowing/lending policies which shareholders may then adopt in the perfect capital market.)
- Show the effects on the firm's investment and production decisions of a tax on operating profit (revenue less variable costs) when
 - interest payments are tax deductible;
 - interest payments are not tax deductible.
- Generalize the model of this section to the case in which the firm is a monopolist in the output market. Will the monopolist's average and marginal rates of return be equal to or exceed r ?
- What restrictions on the production function are necessary to ensure that PP is concave?
- Imperfect capital market.* Analyse the solutions to the optimal investment and consumption decisions when the interest rate at which the owner can borrow differs from that at which she can lend. State the implications for the separation of the two decisions discussed in this section. From that, suggest what difficulties confront a firm in which a salaried manager wishes to take decisions in the interests of n shareholders, where $n \geq 2$.

10. Suppose that the firm could sell its capital equipment at the end of period 1. How would this affect Figs 15.3 and 15.4?
11. *Net present value and internal rate of return.* The net present value of an investment project is

$$NPV(r) = \sum_{t=0}^T \frac{R_t}{(1+r)^t}$$

where R_t is the change in the firm's cash flow in period t caused by the project.

- (a) Show that the firm should accept all projects for which $NPV(r) \geq 0$.
- (b) The *internal rate of return* on an investment project is defined as the rate of interest i at which the net present value is zero: $NPV(i) = 0$. It is often suggested that the rule of accepting all projects for which $i > r$ is equivalent to the NPV criterion. Explain the rationale for this.
- (c) Under what circumstances will the two criteria yield different decisions?
- (Hint: when would $NPV(i) = 0$ have multiple roots? Suppose investment projects are mutually exclusive?)

D. Capital market equilibrium

In the previous sections, we took the market interest rate as given to consumers and firms. However, the interest rate is a price and although the perfect capital market assumption implies that it can be taken as constant by each borrower and lender, its value will be determined by the overall interaction of the decisions of borrowers and lenders. The analysis is not complete until this interaction is examined.

The capital market will be in equilibrium when supply, in the form of lending, equals demand, in the form of borrowing for investment and consumption. To see how this equilibrium is determined we proceed by deriving, from the solutions to the optimization problems of consumers and firms, the relation between the market interest rate and their borrowing or lending. Aggregation of the resulting relationships then leads to the determination of market equilibrium.

We first examine the consumer's demand functions for current M_0^* and future M_1^* consumption which will then yield his net demand for current funds ($M_0^* - \bar{M}_0$) to finance his optimal consumption plans. To simplify the derivation and to emphasize the formal similarities with the analysis of the consumer in Chapter 4, we write the wealth constraint as

$$V_0 = \bar{M}_0 + \mu \bar{M}_1 \geq M_0 + \mu M_1 \quad [\text{D.1}]$$

where $\mu \equiv 1/(1+r)$ is the *discount factor*. We can think of μ as the relative price of period 1 money in terms of period 0 money: £1 of period 1 money can be exchanged for μ of period 0 funds. The Lagrangean for the consumer's problem of maximizing $u(M_0, M_1)$ subject to [D.1] (but assuming the non-negativity constraints on the M_t do not bind) is

$$u(M_0, M_1) + \lambda [\bar{M}_0 + \mu \bar{M}_1 - M_0 - \mu M_1] \quad [\text{D.2}]$$

The resulting first-order conditions determine the consumer's demand for period t consumption

$$M_t^* = M_t^*(\mu, V_0) \quad (t = 0, 1) \quad [\text{D.3}]$$

as functions of the relative price μ and the consumer's wealth V_0 . Substituting these into the direct utility function yields the indirect utility function for the intertemporal consumption problem:

$$u^* = u^*(\mu, V_0) = u(M_0^*, M_1^*) \quad [\text{D.4}]$$

To analyse the effect of changes in μ , and hence in r , on the demand for consumption we use the duality methods of Chapter 4, sections A and B. Consider the problem of minimizing the wealth which must be given to a consumer to enable him to achieve a specified level of utility u :

$$\min_{M_0, M_1} M_0 + \mu M_1 \quad \text{s.t. } u(M_0, M_1) \geq u \quad [\text{D.5}]$$

(ignoring the non-negativity constraints on the M_t). The values of current and future consumption which solve this wealth minimization problem are the Hicksian constant utility demands h_t for consumption as functions of the relative price of period 1 and period 0 consumption and the required level of utility:

$$h_t = h_t(\mu, u) \quad [\text{D.6}]$$

The minimized wealth is

$$m = m(\mu, u) = h_0(\mu, u) + \mu h_1(\mu, u) \quad [\text{D.7}]$$

Applying the Envelope Theorem to [D.7] we have $\partial m / \partial \mu = h_1$, which is a version of Shephard's lemma. The methods of Chapter 4, section A can be used to show that $m(\mu, u)$ is concave in μ so that the own substitution effect on period 1 consumption demand is negative: $\partial h_1 / \partial \mu = \partial^2 m / \partial \mu^2 < 0$. It is also clear from Fig. 15.2 that, since an increase in μ leads to a flattening of the wealth line, the consumer moves around an indifference curve from left to right as μ increases, so that $\partial h_0 / \partial \mu > 0$ and $\partial h_1 / \partial \mu < 0$.

If we set the required utility level in the wealth minimization problem equal to u^* from the consumer's utility maximization problem, then recalling Chapter 4, section B, it must be true that the wealth minimizing Hicksian consumption demands are equal to the utility maximizing Marshallian consumption demands:

$$M_t^*(\mu, V_0) = h_t(\mu, u^*(\mu, V_0)) \quad (t = 0, 1) \quad [\text{D.8}]$$

and the minimized wealth is equal to consumer's wealth in the utility maximization problem: $m(\mu, u^*) = V_0$. Differentiating [D.8] with respect to V_0 gives the effect of increases in wealth on the Marshallian demand

$$\frac{\partial M_t^*}{\partial V_0} = \frac{\partial h_t}{\partial u} \frac{\partial u^*}{\partial V_0} = \frac{\partial h_t}{\partial u} \lambda \quad [\text{D.9}]$$

(where we use the Envelope Theorem on [D.2] to get $\partial u^* / \partial V_0 = \lambda$). Differentiating [D.8] with respect to the relative price μ and using [D.9] gives

$$\begin{aligned} \frac{\partial M_t^*}{\partial \mu} &= \frac{\partial h_t}{\partial \mu} + \frac{\partial h_t}{\partial u} \frac{\partial u^*}{\partial \mu} = \frac{\partial h_t}{\partial \mu} + \frac{\partial h_t}{\partial u} \lambda (\bar{M}_1 - M_1^*) \\ &= \frac{\partial h_t}{\partial \mu} + \frac{\partial M_t^*}{\partial V_0} (\bar{M}_1 - M_1^*) \end{aligned} \quad [\text{D.10}]$$

[D.10] is the Slutsky equation for the intertemporal consumption demand. As in Chapter 4, section 4B, we can decompose the effect of an increase in the relative price μ on the Marshallian demand into a definitely signed substitution effect ($\partial h_0/\partial \mu > 0$ and $\partial h_1/\partial \mu < 0$) and a real wealth effect whose sign is ambiguous. The wealth effect is ambiguous both because the effect of an increase in wealth on demand is ambiguous (consumption in period t may be a normal or an inferior good) and because wealth may be increased or decreased by an increase in μ . For example, if the consumer is borrowing, $\bar{M}_1 - M_1^*$ is positive and an increase in μ makes him better off since he is in effect 'selling' period 1 consumption to finance an increase in period 0 consumption.

To find the effect of an increase in the rate of interest on consumption demands we only need to remember that $\mu = 1/(1+r)$ and to multiply [D.10] by $d\mu/dr = -1/(1+r)^2$ to get $\partial M_1^*/\partial r$ decomposed into a substitution effect and a wealth effect. Since increases in r steepen the wealth line we have $\partial h_0/\partial r < 0$ and $\partial h_1/\partial r > 0$.

In Fig. 15.5 the consumer has an initial endowment (\bar{M}_0, \bar{M}_1) , and we observe how his equilibrium choices vary with changes in the interest rate. The line V_0 in (a) corresponds to interest rate r , and the consumer chooses current consumption of M_0^* . This implies

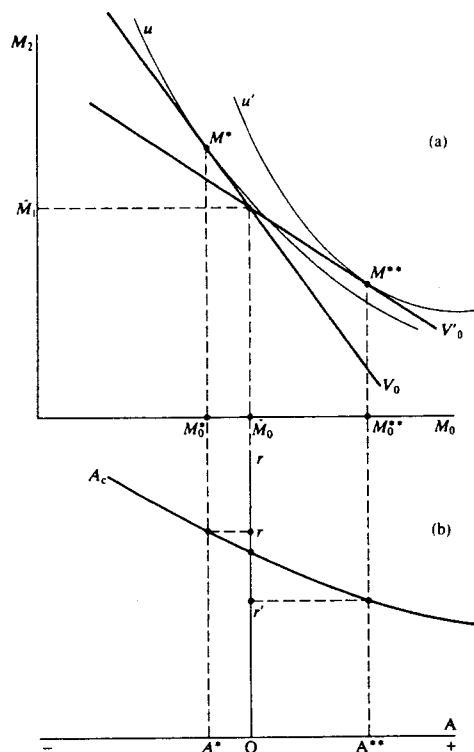


Fig. 15.5

lending $\bar{M}_0 - M_0^*$, shown as $A^* (< 0)$ in (b) of the figure. The line V'_0 corresponds to interest rate r' , and yields an equilibrium current consumption choice at M_0^{**} . This implies borrowing of the amount $M_0^{**} - \bar{M}_0$, shown as $A^{**} (> 0)$ in (b) of the figure. The curve A_c traces out the relation between the market interest rate and the consumer's lending ($A < 0$) or borrowing ($A > 0$). It can be thought of as corresponding to a sequence of equilibrium points in (a) of the figure, such as M^* and M^{**} (compare the analysis of the offer curve in Chapter 3, section E). Fig. 15.5(b) shows an intuitively appealing case: at sufficiently high interest rates, the consumer is a lender. Given his preferences and initial endowment, a falling interest rate causes a decrease in his lending until, after a point at which he neither lends nor borrows, he begins to borrow. Borrowing then varies inversely with the interest rate. In what follows, we shall take the case shown in Fig. 15.5 as typical. (But see question 1, Ex. 15D.)

The effect of changes in r on the intertemporal production and consumption plans of the sole owner of a firm can also be most easily derived by initially using μ as the relative price of period 1 consumption. Recalling the separation theorem from section C, the owner's decisions can be separated into first maximizing wealth via an intertemporal production plan and then maximizing utility subject to the resulting wealth constraint. The constraint on the owner's wealth maximization problem is $D_1 = P(D_0)$ ($P' < 0$, $P'' < 0$) and the wealth maximization problem is to choose D_0 to maximize

$$V = D_0 + \mu D_1 = D_0 + \mu P(D_0) = V(D_0, \mu) \quad [\text{D.11}]$$

and the first- and second-order conditions (assuming a non-corner solution) are

$$\partial V(D_0, \mu)/\partial D_0 = V_D = 1 + \mu P'(D_0) = 0 \quad [\text{D.12}]$$

$$\partial^2 V(D_0, \mu)/\partial D_0^2 = V_{DD} = \mu P'' < 0 \quad [\text{D.13}]$$

These determine the wealth maximizing cash flows from the firm and imply a particular level of capital stock in period 1 and thus a particular level of investment in period 0.

The effect of an increase in μ on the optimal D_0^* is, using the simple comparative statics procedure of Chapter 2, section 2I,

$$\frac{\partial D_0^*}{\partial \mu} = -\frac{\partial^2 V/\partial D_0 \partial \mu}{V_{DD}} = -\frac{\partial[1 + \mu P'(D_0^*)]/\partial \mu}{V_{DD}} = -\frac{P'(D_0^*)}{V_{DD}} < 0 \quad [\text{D.14}]$$

Thus increases in μ reduce the period 0 cash flow from the firm, implying a larger period 1 capital stock and more investment in period 0. Since reductions in r are equivalent to increases in μ , we see that reductions in the rate of interest will reduce the period 0 cash flow and increase period 0 investment and period 1 capital stock. Thus in Fig. 15.6 a reduction in r (or increase in μ) flattens the wealth lines and shifts the optimal production plan from D^* to D^{**} .

The effect of μ on the maximized value V^* of the wealth of the owner is found by using the Envelope Theorem on [D.11] to get

$$dV^*/d\mu = \partial V(D_0^*, \mu)/\partial \mu = V_\mu(D_0^*, \mu) = P(D_0^*) = D_1^* > 0 \quad [\text{D.15}]$$

Hence a reduction in r (which increases μ) leads to an increase in the owner's wealth. This is shown in Fig. 15.6 by the fact that the new wealth line V^{**} has a greater intercept on the horizontal axis than the original wealth line V^* .

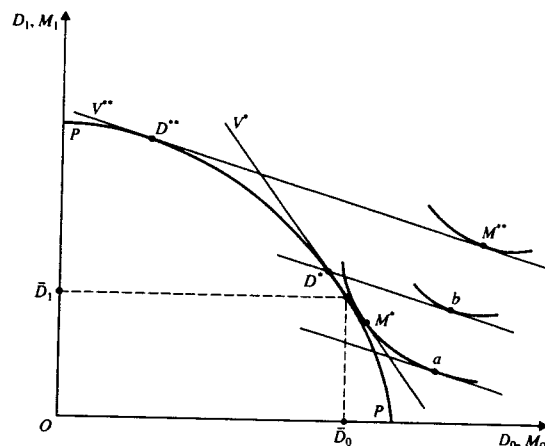


Fig. 15.6

The owner's optimal consumption plan is found by maximizing $u(M_0, M_1)$ subject to the constraint $V^* \geq M_0 + \mu M_1$. The Marshallian period 0 consumption demand is $M_0^*(\mu, V^*)$ and so the effect of an increase in μ is

$$\frac{dM_0^*}{d\mu} = \frac{\partial M_0^*}{\partial \mu} + \frac{\partial M_0^*}{\partial V^*} \frac{dV^*}{d\mu} = \frac{\partial h_0}{\partial \mu} + \frac{\partial M_0^*}{\partial V^*} (D_1^* - M_1^*) + \frac{\partial M_0^*}{\partial V^*} D_1^* \quad [\text{D.16}]$$

where we have used the Slutsky equation [D.10], with D_1^* instead of \bar{M}_1 , to substitute for $\partial M_0^* / \partial \mu$. From [D.16] we see that the effect of an increase in μ on the period 0 consumption of the owner of the firm can be decomposed into three terms. The first term is the substitution effect which is always positive. The last two terms are the two wealth effects of μ . The first of these shows the wealth effect with an unchanged cash flow from the firm and may be positive or negative depending on whether current consumption is a normal or inferior good and whether the owner is borrowing or lending, so that she is made better or worse off by an increase in μ . The second wealth effect arises because the change in μ alters the owner's cash flows from the firm as she changes her wealth maximizing production plan. Her wealth always increases as a result of an increase in μ and so the sign of the second wealth effect depends only on whether period 0 consumption is normal or inferior. We will refer to this effect as the *production effect* on consumption.

Again by using $d\mu/dr = -1/(1+r)^2$ we can translate these results into the effects of an increase in the rate of interest on the owner's consumption plan. In Fig. 15.6 the consumer is initially at M^* and a reduction in r (increase in μ) changes the optimal consumption plan to M^{**} . The substitution effect is from M^* to a , the usual wealth effect is from a to b and the production effect leads to the change from b to M^{**} . The effect of the reduction in r is to increase the firm's investment from $\bar{D}_0 - D_0^*$ to $\bar{D}_0 - D_0^{**}$ and to increase the owner's borrowing from $M_0^* - D_0^*$ to $M_0^{**} - D_0^{**}$. In this case the owner would have a negatively sloped net demand curve for current consumption. As the reader

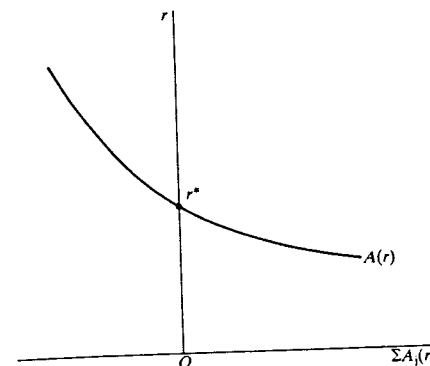


Fig. 15.7

should check, it is easy to construct converse cases if the owner is a lender at sufficiently high interest rates.

When the firm is owned by more than one shareholder the analysis is essentially unchanged: a change in the rate of interest will change the firm's cash flows and the wealth of its owners. Hence there will again be a substitution effect, a wealth effect and a production effect on the lending and borrowing decisions of the individual owners. We assume that the consumers and owner/consumers have well behaved net demand functions for period 0 consumption:

$$A_j = A_j(r) \quad dA_j(r)/dr < 0 \quad j = 1, \dots, J \quad [\text{D.17}]$$

It follows that an equilibrium interest rate r^* satisfies

$$A(r^*) = \sum_{j=1}^J A_j(r^*) = 0 \quad [\text{D.18}]$$

since in that case borrowing ($A_j > 0$) equals lending ($A_j < 0$) in the aggregate, at interest rate r^* . This equilibrium is shown in Fig. 15.7, where $\sum_i A_j = 0$ holds at the interest rate r^* ; the curve $A(r)$ can be regarded as the horizontal sum of all individual curves showing lending and borrowing as functions of the interest rate. (Compare the excess demand functions in Chapters 10 and 16.)

Exercise 15D

1. Construct cases in which:

- a consumer never borrows,
- a consumer never lends,
- the curve A_c in Fig. 15.5(b) has a positive slope over some range.

By separating the effects of an interest rate change into substitution and wealth effects, formulate sufficient conditions under which the curve A_c will always have a negative slope.

2. Analyse the effects on a consumer's borrowing/lending behaviour of:

- (a) a windfall gain in next-period income;
- (b) the imposition of an income tax;
- (c) the imposition of a tax on the returns to lending.

3. Analyse the effects on the market interest rate of:

- (a) an increase in the price of output expected next period;
- (b) an increase in next period's expected wage rate;
- (c) a tax on returns to lending;
- (d) a profits tax (with loan interest not deductible).

4. Analyse the implications for the market interest rate of the changes listed in Question 2 above.

5. Construct a case in which the owner of a firm increases his borrowing following a rise in the interest rate. Explain the relation between the strength of the wealth effect and the curvature of PP in this case.

E. Extensions: many periods; adjustment costs

The analysis of the preceding sections suggested that the equilibrium interest rate, levels of borrowing and lending and investment depend on patterns of consumers' preferences for consumption now as compared to consumption later, future production functions, and expectations about future prices and wage rates. Thus we see that, at least in part, interest and investment are determined by the classical forces of 'productivity and thrift'. However, expectations about future prices and technology also play an important role and to investigate this further let us consider a generalization of the model.

Suppose that there are $T > 2$ periods, indexed $t = 0, 1, \dots, T$. Consider first the consumer who faces a sequence of budget constraints:

$$\sum_{j=1}^m p_{tj} x_{tj} + A_t \leq \sum_{j=1}^m p_{tj} \bar{x}_{tj} + A_{t-1}(1 + r_{t-1}) \quad t = 0, \dots, T \quad [E.1]$$

where p_{tj} is the price of good j expected by the consumer to prevail in year t , x_{tj} is planned consumption of good j in year t , A_t is bond purchase or sale in that year, \bar{x}_{tj} the initial endowment of good j in year t , and r_{t-1} is the interest rate which is expected to prevail in year $t - 1$. Equation [E.1] simply says that in any year, the consumer's expenditure plus net bond purchases cannot exceed the value of endowment of goods plus the net repayment of principal and interest on the consumer's bond purchase of the previous year. The difference from the constraints that were given in equations [B.1] and [B.2] is that now we explicitly incorporate *goods*.

In year 0, $A_{-1}(1 + r_{-1})$ will be a given sum, determined by past decisions, and we denote this by R_0 . In year T , $A_T = 0$, since in effect the economy ceases to exist after that time. Thus we have that in year T :

$$\sum_j p_{Tj}(x_{Tj} - \bar{x}_{Tj})/(1 + r_{T-1}) = A_{T-1} \quad [E.2]$$

from [E.1], where we have dropped the inequality on the premise that a boundary solution will always obtain. Substituting into the budget constraint for year $t - 1$ gives:

$$\sum_j p_{T-1,j} x_{T-1,j} + \sum_j p_{Tj} \frac{(x_{Tj} - \bar{x}_{Tj})}{(1 + r_{T-1})} = \sum_j p_{T-1,j} \bar{x}_{T-1,j} + A_{T-2}(1 + r_{T-2}) \quad [E.3]$$

and so solving for A_{T-2} gives:

$$\sum_j p_{T-1,j} (x_{T-1,j})/(1 + r_{T-2}) + \sum_j p_{Tj} (x_{Tj} - \bar{x}_{Tj})/(1 + r_{T-2})(1 + r_{T-1}) = A_{T-2} \quad [E.4]$$

from which we could then substitute into the budget constraint for year $T - 2$, and so on. Clearly, continuing this process, we would end up with the *single wealth constraint*:

$$\begin{aligned} \sum_j p_{0j}(x_{0j} - \bar{x}_{0j}) + \sum_j p_{1j}(x_{1j} - \bar{x}_{1j})/(1 + r_0) + \dots \\ \dots + \sum_j p_{Tj}(x_{Tj} - \bar{x}_{Tj})/(1 + r_0)(1 + r_1) \dots (1 + r_{T-1}) = R_0 \end{aligned} \quad [E.5]$$

This is the T -period counterpart of the wealth constraint in equation [B.3] earlier, again with the difference that it is written in terms of goods rather than generalized consumption. Now let us define:

$$\begin{aligned} p'_{tj} &= p_{tj}/(1 + r_0)(1 + r_1) \dots (1 + r_{t-1}) \quad t = 1, \dots, T \\ &= p_{tj} \quad t = 0 \end{aligned} \quad [E.6]$$

as the present value, at year 0, of the price of good j in year t . Then [E.5] can be rewritten as:

$$\sum_j \sum_t p'_{tj}(x_{tj} - \bar{x}_{tj}) = R_0 \quad [E.7]$$

Finally, we now define the consumer's utility function on goods rather than general consumption expenditure time-streams, so that:

$$u = u(x) \quad i = 1, 2, \dots, n \quad [E.8]$$

where

$$x = (x_{01}, x_{02}, \dots, x_{Tj})$$

Then we can view the consumer's optimization problem as being to choose values of the goods which maximize u subject to the wealth constraint in [E.7]. Provided the consumer knows the prices p_{tj} and interest rates r_0, \dots, r_{T-1} , this problem is formally no different to that of consumption choices within a time period. Thus, if we characterize a good not only by its physical characteristics, but also by the date at which it is to be consumed, the earlier consumer analysis is directly applicable. In this case the equilibrium of the consumer determines not only a consumption pattern, but also a pattern of lending and borrowing over time. But can we reasonably expect consumers to know future prices and interest rates? We shall consider this question further when we have generalized the model of the firm.

Let y_{tj} be the firm's net output of good j in year t with $y_{tj} < 0$ for an input in year t

and $y_{ij} > 0$ for an output. I_t is the firm's acquisition of the single capital good in year t . K_t is the firm's capital stock in year t and the capital stock is increased by investment and reduced by depreciation according to

$$K_t = K_{t-1}(1 - \gamma) + I_t \quad t = 1, 2, \dots, T \quad [\text{E.9}]$$

$0 \leq \gamma < 1$ is the proportion of capital stock which wears out in one year, i.e. the depreciation rate. The price of a unit of the investment good at time t is p_{tK} and the present value of this price at year 0 is

$$p'_{tK} = p_{tK} / (1 + r_0)(1 + r_1) \dots (1 + r_{t-1}) \quad t = 1, 2, \dots, T - 1 \quad [\text{E.10}]$$

$$p'_{0K} = p_{0K}$$

In each year the firm faces the implicit production function

$$g_t(y_t, K_t) = 0 \quad t = 0, 1, \dots, T \quad [\text{E.11}]$$

(where y_t is the vector of non-capital goods) which defines the feasible input-output combinations in year t . Since the firm operates in competitive input and output markets and takes all prices as given, we assume that diminishing returns set in at a fairly small scale. Its profit in year t is

$$\pi_t = \sum_j p_{tj} y_{tj}$$

(Remember that outputs are measured positively and inputs, apart from capital, negatively.) If the firm is acting in the best interests of its shareholders it will seek to maximize their wealth by maximizing the present value of its cash flow:

$$V = \sum_t (\pi_t - p_{tK} I_t) / (1 + r_0)(1 + r_1) \dots (1 + r_{t-1}) = \sum_t (p'_{tj} y_{tj} - p'_{tK} I_t) \quad [\text{E.12}]$$

[E.12] is the present value of the firm's profits less the present value of its investment expenditures.

The constraints on the maximization of V are the production constraints in each period and the equations [E.9] governing the evolution of the capital stock and the firm's initial capital stock K_0 . The Lagrangean for the problem is

$$\mathcal{L} = V + \sum_t \beta_t g_t(y_t, K_t) + \sum_t \alpha_t [K_{t-1}(1 - \gamma) + I_{t-1} - K_t] \quad [\text{E.13}]$$

and the first-order conditions are

$$\partial \mathcal{L} / \partial y_{tj} = p'_{tj} + \beta_t g_{tj} = 0 \quad t = 0, 1, \dots, T \quad j = 1, \dots, J \quad [\text{E.14}]$$

$$\partial \mathcal{L} / \partial K_t = \beta_t g_{tK} - \alpha_t + \alpha_{t+1}(1 - \gamma) = 0 \quad t = 1, 2, \dots, T \quad [\text{E.15}]$$

$$\partial \mathcal{L} / \partial I_t = -p'_{tK} + \alpha_{t+1} = 0 \quad t = 0, 1, \dots, T - 1 \quad [\text{E.16}]$$

where $g_{tj} = \partial g_t / \partial y_{tj}$ and $g_{tK} = \partial g_t / \partial K_t$. (Notice that since there is no production after period T there is no point in making any investment in period T .) These conditions together with the constraints determine the firm's production plan from year 0 to year T , including its purchase of investment goods and the evolution of its capital stock. We can extract some insights from the conditions by recalling from Chapter 7, section D, that we can

interpret $-g_{tK} / g_{tj} = \partial y_{tj} / \partial K_t$ as the marginal product MP_{Kjt} of capital in the production of good j when $y_{tj} > 0$, or as the marginal rate of technical substitution $MRTS_{Kjt}$ between capital and input j if $y_{tj} < 0$. If we take the condition on good j at time t and write it as $\beta_t = -p'_{tj} / g_{tj}$ and use [E.16] to substitute for α_t in [E.15] we get, (assuming for definiteness that good j is an output at time t):

$$p'_{tj} MP_{Kjt} = p'_{t-1K} - p'_{tK}(1 - \gamma) \quad [\text{E.17}]$$

Recalling the definition of the present value prices p'_{tj} and p'_{tK} from [E.6] and [E.10], we can multiply both sides of [E.17] by $(1 + r_0)(1 + r_1) \dots (1 + r_{t-1})$ to get

$$p_{tj} MP_{Kjt} = (1 - r_{t-1}) p_{t-1K} - p_{tK}(1 - \gamma) \quad [\text{E.18}]$$

If we define the proportionate rate of growth θ in the price of the investment good between period $t - 1$ and period t we can write p_{tK} as $p_{t-1K}(1 + \theta)$ and rearrange [E.18]

$$p_{tj} MP_{Kjt} = p_{t-1K}(r_{t-1} + \gamma - \theta + \rho\gamma) \quad [\text{E.19}]$$

If we finally assume that the rate of interest is constant and that γ and θ are both relatively small we have

$$p_{tj} MP_{Kjt} \approx p_{t-1K}(r + \gamma - \theta) \quad [\text{E.20}]$$

The left-hand side of [E.20] is the value of the marginal product of capital: the value of the extra output produced in period t by having an extra unit of capital at that date. (Notice that the value of the marginal product of capital is the same whichever output j is considered.) The right-hand side is the rental price of capital in period t . To acquire the use of an extra unit of capital for period t the firm can be thought of as buying an extra unit of the investment good in year $t - 1$ and then reselling what is left of the unit after one period. The cost to the firm of these transactions depends on (a) the interest r on the foregone cash flow used to purchase the investment good; (b) the loss of γ of the extra unit of the investment good from depreciation; and (c) the gain or loss to the firm from the change in the price of the investment good at the rate θ between the period $t - 1$ when the extra unit is bought and period t when it is resold.

The condition [E.20] is thus a generalization of the results in section C to allow for many goods, many periods and depreciation. We leave it to the reader to examine the case in which good j is an input by substituting $MRTS_{Kjt}$ for MP_{Kjt} in [E.20] and then dividing through to get the generalization of the minimum cost production condition that the marginal rate of technical substitution should be equal to the marginal costs of the inputs to the firm.

This discussion serves to emphasize the importance of the expectations of consumers and firms. We can envisage an extreme case, in which all consumers and firms expect the same prices, and at these prices, all planned supplies and demands are consistent, so that the expected prices are the true equilibrium prices. In that case, all consumers' lending/borrowing and consumption plans will actually be realized, as will firms' production and investment plans. Equivalently, we could imagine that at year 0, markets are held in which are exchanged claims to specified goods at each future date, and claims to wealth at each date. In other words, there would be a market for every j and every t , held at year 0. In that case, the prices p'_j would actually be established at that date, and consumers and firms

would then spend the rest of time ($t = 1, 2, \dots$) honouring the commitments they made at year 0.

Neither of these cases appears to be an adequate description of the real world. Although some futures markets exist, most goods are traded on 'spot' markets, i.e. markets are held at every time t , including capital markets. Moreover, at any time, consumers may not know with certainty the tastes they will have at some future time, nor what their endowments of wealth will be. Similarly, firms may not know with certainty future technological possibilities. Expectations may differ about future prices and interest rates and the plans made by consumers and firms may be inconsistent, and so a given consumption or production plan may not be capable of realization at a given date. Firms and consumers are likely to be aware of this, and that awareness may influence the decisions they take at any one time. Thus, decisions over time can be handled without any change in the formal structure of analysis if we rule out uncertainty about future prices, tastes, and technology. But this means ignoring what appears to be an important and pervasive aspect of economic activity. Hence, in Chapters 19 to 22, we consider some elements of the economics of uncertainty.

Adjustment costs

In our analysis of the firm in Chapters 7 to 9 we made a distinction between the long run and the short run by defining the short run as a length of time within which the firm is unable to vary all of its inputs. This distinction is useful but it is a crude attempt to capture the effects of adjustment costs. These are the costs which arise solely from a *change* in the level of use of an input. For example if a firm wishes to hire more labour it may have to advertise for new workers, interview and otherwise screen them to see if they are suitable, and train them. Such costs are adjustment costs: they are incurred solely because the firm wishes to hire more workers. The firm's costs in any period depend both on the number of workers employed and on the increase in the number of workers. Similarly, if the firm wishes to increase its capital stock it may find that the price of capital goods or the costs of installing them increase with the level of investment.

Adjustment costs explain why firms do not immediately change the levels of input use in response to changes in market conditions. Since adjustment costs associated with different types of input are likely to differ, they may also explain why some types of input are varied more frequently than others. Thus we can regard the analysis earlier in this chapter, in which we assumed that labour input can be varied within a period but the capital stock cannot, as resting on tacit assumptions about the nature of adjustment costs. Labour was assumed to have no adjustment costs and therefore was the variable factor, whereas it was tacitly assumed that the cost of adjusting the capital stock within a period was extremely high, so that a rational firm would not *wish* to change its capital stock within a period but would change output solely by changing the amount of labour employed.

To show the implications of adjustment costs for the behaviour of the firm we will extend the two period model with no adjustment costs set out in section C. The firm operates in a stationary environment: the prices of output p , labour w and investment goods p_K , the interest rate r , the depreciation rate γ , and the firm's strictly concave production function $y_t = f(L_t, K_t)$ are the same in all periods.

We assume that the firm's expenditure on capital goods in any period is given by

$$c(I_t) = p_K I_t + \frac{1}{2} a I_t^2 \quad [\text{E.21}]$$

where $a > 0$ is an adjustment cost parameter. One rationalization for this cost of investment function might be that the firm must incur costs to install new equipment and that these costs increase more than in proportion to the amount of new capital installed. Other plausible specifications might have the price of the capital good increasing with the amount bought in any period. The particular form of $c(I_t)$ will vary with the circumstances but what is important for the conclusions to be drawn is that $c(I_t)$ is a convex function of investment: $c'' > 0$. (See Question 3, Exercise 15E.)

The firm chooses its inputs and investment to maximize its discounted present value

$$V = \sum_t [pf(L_t, K_t) - wL_t - p_K I_t - \frac{1}{2} a I_t^2] / (1+r)^t \quad [\text{E.22}]$$

subject to the constraint linking the capital stock in successive periods:

$$K_t = K_{t-1}(1 - \gamma) + I_{t-1} \quad [\text{E.23}]$$

and to the firm's initial capital stock K_0 . We assume that the firm has an infinite life. The Lagrangean for the problem is

$$\mathcal{L} = V + \sum_t \alpha_t [K_{t-1}(1 - \gamma) + I_{t-1} - K_t] \quad [\text{E.24}]$$

and the first-order conditions are

$$\partial \mathcal{L} / \partial L_t = [pf_{L_t} - w] / (1+r)^t = 0 \quad t = 0, 1, \dots \quad [\text{E.25}]$$

$$\partial \mathcal{L} / \partial K_t = pf_{K_t} / (1+r)^t - \alpha_t + \alpha_{t+1}(1 - \gamma) = 0 \quad t = 1, 2, \dots \quad [\text{E.26}]$$

$$\partial \mathcal{L} / \partial I_t = -[p_K + a I_t] / (1+r)^t + \alpha_{t+1} = 0 \quad t = 0, 1, \dots \quad [\text{E.27}]$$

where f_{L_t} and f_{K_t} are the marginal products of labour and capital in period t .

In each period the firm chooses its employment (and therefore its output) and its investment, which determines its capital stock for the following period. From [E.25] we have $pf_{L_t} = w$, which is the usual profit maximization requirement that labour should be employed up to the point at which the value of its marginal product is equal to the wage rate. The firm's short-run marginal cost in period t is $SMC(y_t, K_t) = w/f_{L_t}(K_t, L_t)$ which is the cost of increasing output when the capital stock is fixed. In each period the firm maximizes its profit in that period by choosing the output y_t at which the period's price is equal to short-run marginal cost:

$$p = w/f_{L_t}(K_t, L_t) = SMC(y_t, K_t)$$

Thus output and the labour input are not directly affected by adjustment costs associated with the capital stock. However, since the marginal product of labour $f_{L_t}(L_t, K_t)$ and thus the short-run marginal cost, depend on the size of the capital stock, employment and output are indirectly influenced by adjustment costs if these affect the capital stock at any date.

Using [E.27] to substitute for α_t and α_{t+1} in [E.26], then using [E.23] to substitute

$K_{t+1} - (1 - \gamma)K_t$ for I_t and multiplying through by $(1 + r)^t$ we have

$$pf_{K_t}(L_t, K_t) = p_K(r + \gamma) + a\{K_t[1 + r + (1 - \gamma)^2] - (1 - \gamma) \times [(1 + r)K_{t-1} + K_{t+1}]\}$$

$$t = 1, 2, \dots \quad [\text{E.28}]$$

which yields the time path of the firm's capital stock (and hence its investment demand in each period).

We can bring out the implications of adjustment costs for the firm's capital stock by first considering the case in which there are no adjustment costs: $a = 0$. Then [E.28] reduces to

$$pf_{K_t}(L^*, K^*) = p_K(r + \gamma) \quad t = 1, 2, \dots \quad [\text{E.29}]$$

which requires the firm to choose its capital stock in periods 1, 2, ... so as to equate the value of the marginal product of capital to the rental price of capital. Recalling the definition of long-run marginal cost from Chapter 8, section 8B, we see that [E.28] and [E.29] imply that in periods 1, 2, ... the firm plans to produce where price equals long-run marginal cost:

$$p_K(r + \gamma)/f_K(K^*, L^*) = w/f_L(K^*, L^*) = LMC = p$$

Because there are no adjustment costs, the firm chooses its next period capital stock so that LMC is equal to its anticipated next period price. In the next period its actual output is determined by equating SMC to the actual price. When the firm's price expectations are correct its capital stock chosen in the previous period will be such that $p = SMC = LMC$. Because the firm operates in a stationary environment, its price expectations are correct and there are no adjustment costs associated with labour or capital, the firm chooses the same capital stock K^* and labour input L^* in all periods after the initial period 0. In the absence of adjustment costs the firm chooses a level of investment in period 0 which enables it to jump straight to the long-run optimal capital stock K^* in period 1: $I_0 = K^* - (1 - \gamma)K_0$. Thereafter, its investment is just sufficient to offset the depreciation on the capital stock: $I_t = \gamma K^*$ ($t = 1, 2, \dots$).

In Figs 15.8 and 15.9 the firm has inherited a capital stock of K_0 which determines its short-run marginal cost curve $SMC(y, K_0)$ for period 0 and it will choose a period 0 output of y_0 where $p = SMC(y_0, K_0)$. Given its LMC curve and anticipated period 1 price of p it plans to produce y^* in period 1 where $p = LMC = SMC(y^*, K^*)$. Hence its required period 0 investment is shown in part (b) of Fig. 15.8 as $I_0 = K^* - K_0(1 - \gamma)$. In period 1 and subsequently the firm finds that it has correctly anticipated the output price to be p and it chooses the output y^* where $p = SMC(y^*, K^*)$. Its period 1, 2, ... investment is just sufficient to keep the capital stock at the new long-run equilibrium: $I_t = \gamma K^*$. Thus in the absence of adjustment costs and given its correct expectations, the firm jumps straight to the new equilibrium optimal capital stock in period 1 and maintains this stock until there is a change in its environment.

Now consider the long-run optimal capital stock when there are adjustment costs ($a > 0$). The firm will be in long-run equilibrium when its optimal capital stock is the same in

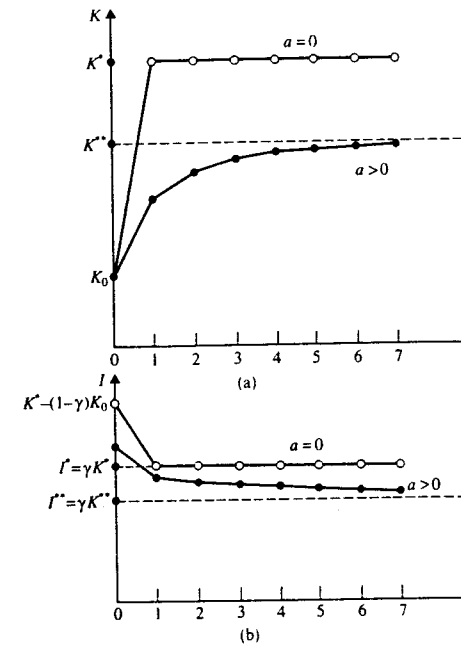


Fig. 15.8

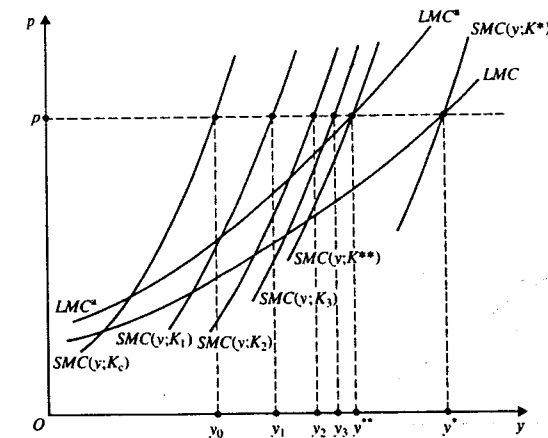


Fig. 15.9

period $t - 1, t, t + 1, \dots$. From [E.28] this implies a capital stock K^{**} which satisfies

$$\begin{aligned} pf_K(L^{**}, K^{**}) &= p_K(r + \gamma) + aK^{**}\{[1 + r(1 - \gamma)^2] - (1 - \gamma)[(1 + r) + 1]\} \\ &= p_K(r + \gamma) + aK^{**}\gamma(r + \gamma) = (p_K + a\gamma K^{**})(r + \gamma) \\ &= (p_K + aI^{**})(r + \gamma) \end{aligned} \quad [E.30]$$

If we assume that the production function is homothetic (see Chapter 7, section 7B), the long-run equilibrium cost minimizing capital-labour ratio is smaller than in the case with no adjustment costs. The fact that the production function is homothetic means that the slope of the firm's isoquants is determined only by the capital-labour ratio and is not affected by its output level. The slope of the firm's isocost curves reflects the relative cost of capital and labour and for cost minimization the firm must choose an input combination where its isoquant and isocost curve are tangent. In long-run equilibrium the firm invests $I^{**} = \gamma K^{**}$ each period to keep the capital stock constant. The cost per period of a marginal increase in the long-run equilibrium flow of capital services exceeds $p_K(r + \gamma)$ because the increased capital stock implies additional replacement investment which increases adjustment costs by $a\gamma K^{**} = aI^{**}$ each period. The marginal cost of capital services in long-run equilibrium is therefore $(p_K + a\gamma K^{**})(r + \gamma)$, which not only exceeds $p_K(r + \gamma)$, but also increases with the long-run equilibrium capital stock. The isocost loci showing long-run input combinations with the same cost are curves, like A in Fig. 15.10, which have a slope $-w/(p_K + a\gamma K^{**})(r + \gamma) > -w/p_K(r + \gamma)$. A becomes steeper at smaller capital-labour ratios and the long-run equilibrium capital-labour combination is at K^{**}, L^{**} on the isoquant y^{**} corresponding to the firm's optimal long-run output.

If there were no adjustment costs ($a = 0$) the firm's isocost curves would be straight lines with slope $-w/p_K(r + \gamma)$. The long-run equilibrium input mix would then be at K^*, L^* where an isocost line with slope $-w/p_K(r + \gamma)$ is tangent to the isoquant for the firm's long run optimal output y^* .

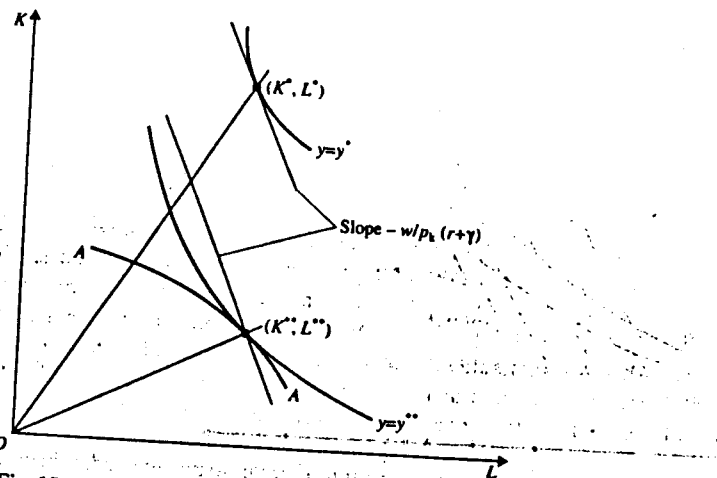


Fig. 15.10

Note that adjustment costs reduce the firm's long-run equilibrium output: $y^* > y^{**}$. With adjustment costs the long-run marginal cost of the firm is

$$(p_K + a\gamma K^{**})(r + \gamma)/f_K(K^{**}, L^{**}) = w/f_L(K^{**}, L^{**}) = LMC^a$$

which increases with the firm's adjustment cost parameter a . Since, from [E.30], the firm's long-run equilibrium output satisfies $p = LMC^a$, we see that adjustment costs reduce the firm's long-run equilibrium output as well as reducing its long run equilibrium capital-labour ratio.

Comparing [E.29] and [E.30] it is apparent that adjustment costs affect the long-run equilibrium of the firm in this model *only* if the depreciation rate is positive. When $\gamma = 0$ the firm does not need to make replacement investment to maintain the capital stock at its long run equilibrium level and so does not incur adjustment costs in long-run equilibrium. However, adjustment costs do affect the time path of output and input choices. The firm will not move to the long-run equilibrium in period 1 if it faces adjustment costs: in each period it will partially adjust its capital stock towards the long-run equilibrium.

Figs 15.8 and 15.9 can be used to show the firm's responses to an unanticipated increase in the price of its output when it has adjustment costs. Suppose that it has an initial capital stock of K_0 giving it the short-run marginal cost curve $SMC(y, K_0)$ in period 0. It will produce the output y_0 where the period 0 price equals its period 0 short-run marginal cost. It will plan to increase its capital stock to K_1 by its period 0 investment of I_0 . In period 1 the firm will have the capital stock K_1 and thus the short-run marginal cost curve $SMC(y, K_1)$. It will maximize period 1 profit by producing y_1 where the period 1 price p equals its short-run marginal cost. It will again plan to increase its capital stock K_2 by investing I_1 in period 1. In period 2 the firm will have the short-run marginal cost curve $SMC(y, K_2)$ and will choose an output of y_2 in period 2 where the period 2 price p equals marginal cost. It will increase its period 3 capital stock by its investment I_2 . Eventually the firm will attain its desired long-run equilibrium capital stock K^{**} and produce the long-run equilibrium output y^{**} where $p = LMC^a = SMC(y^{**}, K^{**})$.

As is the case when there are no adjustment costs, the firm's long-run equilibrium supply is determined by its long-run marginal cost curve and its short-run decisions in any period by its short-run marginal cost curve. Adjustment costs affect the firm's long supply because they shift the firm's long-run marginal cost curve upward, so its long-run supply is smaller than if there are no adjustment costs. Adjustment costs also mean that the firm will take longer to reach its new long-run equilibrium if there is any change in its environment. The firm's short-run supply responses will be indirectly affected by adjustment costs because they alter the time path of its capital stock and thus the position of its short-run marginal cost curve in a period.

The Marshallian short run-long run distinction rests on a tacit assumption that the firm has infinite adjustment costs of capital for the current period and then zero adjustment costs in the next period. In our treatment we have assumed that adjustment costs are infinite in the current period but positive and finite in the next period. This formulation is still somewhat special but it does bring out some of the implications of adjustment costs. A more general model of adjustment by the firm would recognize that the costs of increasing its capital stock may depend not just on how large the increase is but also on how far in advance the increase is planned. For example, it is not implausible that a given increase in K is less expensive if it is decided two periods beforehand rather than one. We have

also assumed that the firm faces the same price for capital goods whether it buys or sells them and that adjustment costs are symmetric: the cost of reducing the capital stock by a given amount is the same as increasing it by that amount. A proper treatment of the interesting issues which arise if these simplifying assumptions are dropped requires rather more advanced techniques than we use in this book and readers who wish to pursue them should consult the references.

Exercise 15E

1. *Time inconsistency.* At date 0 an individual has preferences represented by the utility function $u^0(c_0, c_1, \dots, c_T)$ where c_t is consumption expenditure at date t . At date i her preferences are represented by the utility function $u^i(c_i, c_{i+1}, \dots, c_T)$. At date 0 the individual chooses the optimal consumption plan $c_0^0, c_1^0, \dots, c_T^0$ which maximizes her utility u^0 subject to the date 0 wealth constraint

$$\sum_{t=0}^T \frac{c_t}{(1+r)^t} \leq V_0$$

where V_0 is the value of her initial wealth. At date i the individual will choose the optimal consumption plan $c_i^i, c_{i+1}^i, \dots, c_T^i$ to maximize u^i subject to the date i wealth constraint

$$\sum_{j=0}^{T-i} \frac{c_{i+j}}{(1+r)^j} \leq V_i = \sum_{j=0}^{T-i} \frac{c_{i+j}^0}{(1+r)^j}$$

where V_i is her period i wealth implied by her date 0 optimal consumption plan. The consumer exhibits time inconsistency if her optimal plan changes over time, i.e. if $c_{i+j}^0 \neq c_{i+j}^i$. What must be assumed about the consumer's intertemporal preferences, as represented by u^0, u^i , to ensure that she will not wish to change her plans at any date i ? (*Hint:* since the relative prices of consumption at different dates in the future, as implied by interest rates, do not alter as time elapses, what must be true of the marginal rates of substitution if plans are not to change? What preference structure does this remind you of?)

2. Suppose that the firm has monopoly power in one of its output markets. How would this affect time path of its capital stock?
3. What would be the effect of assuming that adjustment costs were linear in I_t : $c(I_t) = aI_t$?

References and further reading

Intertemporal resource allocation is considered in some depth in:

- J. Hirshleifer. *Investment, Interest and Capital*, Prentice-Hall, Englewood Cliffs, NJ, 1970.
C. J. Bliss. *Capital Theory and the Distribution of Income*, North Holland, Amsterdam, 1975.

Dynamic optimization is an essential tool for examining many interesting intertemporal problems. A good introduction is:

- A. Dixit. *Optimization in Economic Theory*, Oxford University Press, 2nd edn, 1990, chs 10, 11.

whilst:

N. L. Stokey, R. E. Lucas with E. C. Prescott. *Recursive Methods in Economic Dynamics*, Harvard University Press, Cambridge, MA, 1989.

provides an exhaustive but much more advanced account.

A detailed analysis of the implications of adjustment costs is:

R. E. Lucas. 'Adjustment costs and the theory of supply', *Journal of Political Economy*, 75, 1967, 321-34.